

UNCLASSIFIED

AD 295 571

*Reproduced
by the*

**ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA**

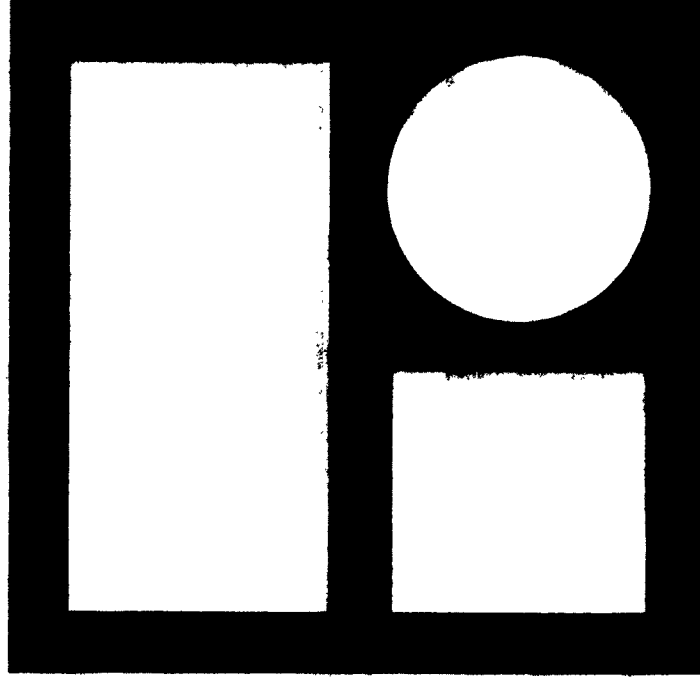


UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

ASTIA

295571



63-2-3

295 571

AFCRL - 62-946

DATA SYSTEMS DIVISION
LITTON SYSTEMS, INC.
221 Crescent Street
Waltham 54, Massachusetts

Contract No. AF19(604)-8828

Project No. 4610

Task No. 461002

Prepared for:

ELECTRONICS RESEARCH DIRECTORATE
Air Force Cambridge Research Laboratories
Office of Aerospace Research
UNITED STATES AIR FORCE
Bedford, Massachusetts

Period Covered:

15 July 1961 - 14 October 1962

Final Report

Investigation of Automation of Speech Processing
for Voice Communication

By George S. Sebestyen/David Van Meter

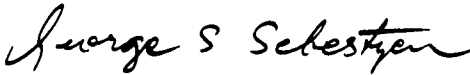
**Investigation of Automation of
Speech Processing for Voice Communication**

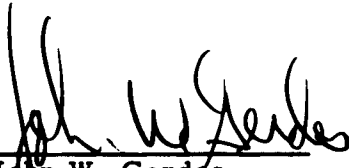
Contract AF19(604)-8828

15 June 1961 to 14 October 1962

**Prepared by
George S. Sebestyen
David Van Meter**

Approved by


**George S. Sebestyen
Technical Director**


**John W. Gerdes
Assistant Manager**

**LITTON SYSTEMS, INC.
DATA SYSTEMS DIVISION**

"Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to the:

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

Department of Defense contractors must be established for ASTIA services or have their 'need-to-know' certified by the cognizant military agency of their project or contract."

"All other persons and organizations should apply to the:

U. S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D. C."

ABSTRACT

This report discusses analytic and experimental procedures applicable to improving the performance of a particular type of speech compression system presently being developed at AFCRL under the direction of C. P. Smith.

The system operates on the principle that the number of linguistically distinguishable varieties of instantaneous spectrum patterns in speech is much smaller than the total number used during any utterance. Considerable compression can be achieved by replacing original speech patterns by reference patterns selected from a relatively small library, and transmitting correspondingly short index numbers or descriptors identifying the reference patterns so used.

The main problem dealt with in the present study concerns the choice of patterns to be used in the reference library. The minimum number of references needed and their exact specification depend on characteristics of human speech that can be determined only by experiment. Material contained in the report relevant to the problem of choosing references includes the following:

a) Formulation of experimental objectives in terms of possible distributions of raw speech patterns and their dependence on the input state, i. e., on speaker, duration of utterance and text material.

b) A technique for library comparison based on minimum absolute distance. This is applied in designing experiments to evaluate the analyzer sampling error and adjust the synthesizer channel gains. It also furnishes an index to which analysis of variance techniques may be applied in order to determine the variability of reference subsets with speaker, duration of utterance, and type of text material.

c) An approximate model for studying the rate of reference generation. This model is based on classical occupancy theory, and with the addition of a time varying parameter may lead to a better understanding of the process of reference generation. Experimental results and their interpretation from this point of view are given.

d) A spherical clustering technique which may be used to reduce a given library to one of smaller size in such a way that each pattern in the original library is within a prescribed absolute distance of some pattern in the reduced library. This technique is applicable to the problem of finding the maximum average radius of exchange subsets, i. e. , subsets within which raw speech patterns may be exchanged freely without serious degradation of speech quality. Examples of its use on actual data are given.

e) Experimental designs for determining whether significant differences in reference libraries may be attributed to different speakers, differences in duration of utterance or differences in text material.

Brief attention is also given to the speech segmentation problem and to encoding techniques. The report concludes with a summary of the program of experimentation recommended for continued investigation.

ACKNOWLEDGMENT

It is a pleasure to acknowledge the encouragement and assistance given by C. P. Smith of the AFCRL Communication Sciences Laboratory, who contributed to the development of the concepts and procedures presented here through informal discussion and constructive criticism, and provided speech data in digitized vocoder format for the illustrative examples.

TABLE OF CONTENTS

ABSTRACT	iv
ACKNOWLEDGMENT	vi
1. INTRODUCTION	1
2. LINGUISTIC EXCHANGE SUBSETS	6
2.1 Subset Configurations	6
2.2 System Implications	12
2.3 Comparison of Reference Libraries	16
3. DETERMINATION OF EQUIPMENT ERRORS	24
3.1 Description of the Voice Data Processing Equipment	24
3.2 Characterization of Sampling Errors	27
3.3 Minimization of Synthesizer Errors	30
4. RATE OF REFERENCE GENERATION	31
4.1 Occupancy Theory Model	31
4.2 Illustrative Examples	32
5. A LIBRARY REDUCTION TECHNIQUE	38
5.1 Spherical Clustering	38
5.2 Illustrative Examples	39
5.3 Machine Computation	50
6. LIBRARY VARIABILITY	55
6.1 Analysis of Variance	55
6.2 Variations with Duration of Utterance	61
6.3 Variations Among Speakers	66
6.4 Variations with Type of Text	67
7. TECHNIQUES FOR FURTHER IMPROVEMENT OF COMPRESSION RATIO	68
7.1 Coding	68
7.2 Segmentation	72
8. RECOMMENDED PROGRAM OF EXPERIMENTATION	77
8.1 Determination of Equipment Errors	77
8.2 Average Diameter and Spacing of Subsets	77

TABLE OF CONTENTS (cont.)

8.3 Rate of Subset Generation	78
8.4 Variations With Time of Speaking	78
8.5 Variations Among Speakers	78
8.6 Variations Due to Type of Text	79
8.7 Speech Segmentation	79
8.8 Statistical Properties of Reference Sequences	79

REFERENCES

LIST OF FIGURES

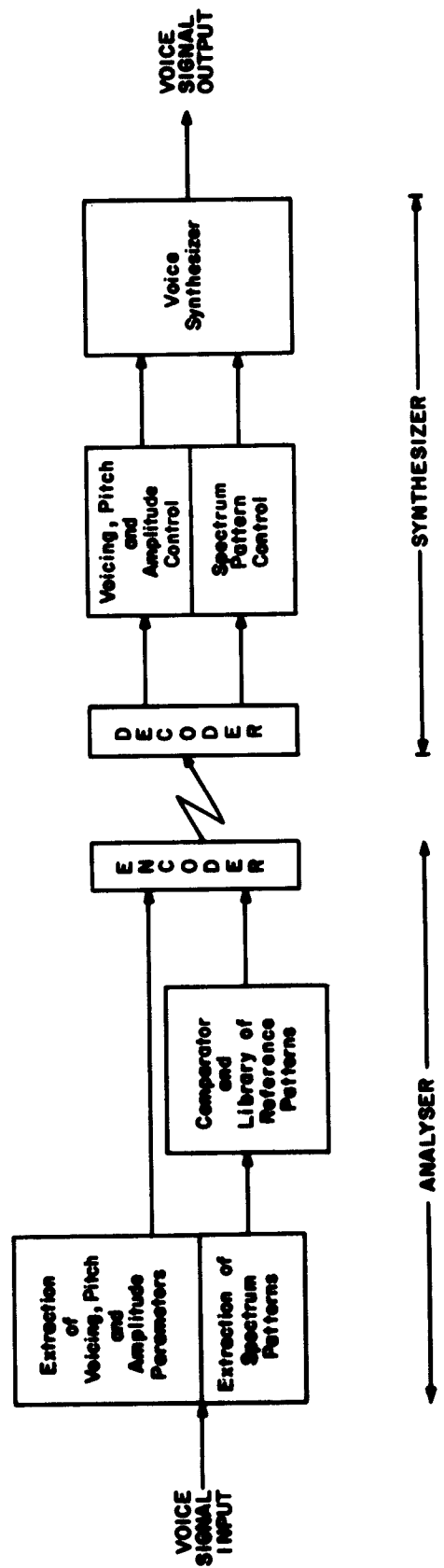
<u>Figure No.</u>	<u>Description</u>	<u>Page No.</u>
1.1	Speech Compression System	2
2.1	Linguistic Subsets vs Machine Subsets	13
2.2	Linguistic Subsets vs Machine Subsets	14
2.3	Twenty Most Frequently Occurring References $T_M = 5$ (CPS-29 August 1961)	18
2.4	Absolute Distances Between References	19
4.1	Rate of Reference Generation $T_M = 5$, 2-Bit Quantization	33
4.2	Rate of Reference Generation $T_M = 3$, 2-Bit Quantization	35
5.1	Histograms of Absolute Distances	40
5.2	Distances after Thresholding ($S=10$)	41
5.3	Choice of New References ($S=10$)	43
5.4	Distances Between References ($S=10$)	44
5.5	Vector X is Close to 2 But Would be Classified Differently	45
5.6	Choice of New References ($S=8$)	48
5.7	Distances Between References ($S=8$)	49
5.8	Machine Clustering	52, 53, 54
6.1	Sample Size vs Number of Families	60
7.1	Speech Segmentation	75
Flow Chart		
2.1	Computation of Distance Between Libraries	23
5.1	Spherical Clustering Computation	51

I. INTRODUCTION

The work described herein is directed toward the acquisition of fundamental knowledge about speech necessary to assess the ultimate capabilities of a particular speech compression technique presently being developed at AFCL under the direction of C. P. Smith. The emphasis is on analytic and experimental techniques that can be used with the AFCL Voice Data Processing System (VDPS), which was designed specifically as a flexible research tool and system simulator for studies of this technique. (1) (4) (5) (6)

Figure 1.1 shows the operations performed by the analyzer and synthesizer of the type of system under investigation. (1) A digital representation of the voice signal input is obtained at the analyzer, consisting of a voiced-unvoiced indication (1 bit), pitch frequency (3 bits), voice amplitude (6 bits), and spectrum pattern (54 bits). The latter is obtained by quantizing the outputs of an 18-channel filter bank (Vocoder) with three-bit resolution per channel. (2) (3) (One or two bit resolution may be used alternatively.) These measurements are repeated every 20 ms, so that information about the incoming speech is generated at a rate of $50 \times (54 + 6 + 3 + 1) = 3200$ bits per second.

Rather than transmitting the spectrum patterns as measured at the vocoder output, the analyzer instead compares each measured pattern with each of a number of reference patterns previously determined and maintained in storage. The number of stored reference patterns is much smaller than the number of varieties of input pattern possibilities. Therefore by selecting the reference pattern which most nearly matches the incoming "raw speech" pattern, the analyzer restricts the number of different pattern "messages" that need be transmitted. Of course, the pattern configurations themselves (54 bits) are not transmitted; instead, a set of code words is used which represents the set of reference patterns with as few bits per word as possible.



WF 68-134

FIGURE 1. Speech Compression System

Thus, by minimizing the number of references used and the lengths of the code words which represent them, a considerable reduction in the data rate of the system is expected. If, for example, 1000 references were found sufficient, not more than 10 bits per pattern on the average would be necessary to represent them without ambiguity. At the sampling rate of 50 per second, the analyzer would then transmit at a rate of only $50 \times (10 + 6 + 3 + 1) = 1000$ bits per second, thus permitting a considerable reduction of bandwidth. The synthesizer receives the incoming digital signal corrupted by noise and decodes it, correcting errors caused by the noise (to an extent determined by the nature of the code used for representing signals at the analyzer output), and separating the parts of the digital sequence corresponding to the spectrum pattern and the voicing, amplitude and pitch parameters respectively. The latter are used to control the excitation of a bank of synthesizer filter-amplifiers, whose relative gains are set in accordance with the received spectrum pattern information.⁽⁵⁾

For a system of this type it is clear that the crucial design factor is the choice of the spectrum patterns that are to be stored and used for references at the analyzer. To obtain a high degree of speech compression we want to use the smallest possible number of references. On the other hand, the number must be sufficiently large, and their locations in 18-dimensional space must be so chosen, that any raw speech pattern input is close enough to some one of the references that the two can be exchanged without serious degradation of speech quality.*

*The possibility of such exchanges may be expected to depend strongly on context, the more so as the distance between the two patterns exchanged increases. On the other hand, we may hope that for sufficiently close patterns this dependence may be ignored without serious effect. The degree of compression obtainable, however, depends on how large this common threshold turns out to be.

Moreover, we should like to operate with a fixed set of references that is independent of the identity of the speaker, the text he speaks or the duration of his utterance. The feasibility of this can only be assessed when we know:

- a) how many different patterns are used by a typical speaker,
- b) how close a pattern must be to a reference before an exchange can be tolerated,
- c) what types of variability occur from speaker to speaker, text to text, and time to time.

Most of the present report is devoted to various aspects of the problem of choosing references. A sample space model to help visualize the variety of system possibilities is discussed in Section 2. Experimental and analytical techniques for studying how the number of references needed by a speaker varies with length of utterance, and for systematically reducing a preliminary library of stored references, are discussed in Sections 4 and 5 respectively. Section 6 considers techniques for studying the variability of the reference libraries used by different speakers at different times with different text materials.

Another important problem concerns the encoding of the reference patterns selected for transmission. It is well known that if advantage can be taken of the statistical dependence among successively occurring reference patterns, a considerable reduction of the average code word length, and hence of the data rate, can be achieved. Section 7.1 discusses this problem.

In Section 7.2 we describe a technique for segmenting speech into intervals, throughout each of which a single reference pattern may be used repetitively without excessive degradation of the quality of the result. The preliminary results given suggest that this technique may be developed to yield considerable compression with rather simple implementation.

Some consideration is given to the determination of equipment errors and the optimization of synthesizer channel gains in Section 3. Problems of voiced-unvoiced switching and the control of the pitch source are not considered here, however; the emphasis throughout is on optimizing the processing of the vocoder spectrum pattern parameters.

2. LINGUISTIC EXCHANGE SUBSETS

2.1 Subset Configurations

Before experiments can be systematically planned, we must be clear on just what revelations are expected from the experimental data, and what relation they have to the overall needs of the program. For this purpose it is necessary to have a specific model in mind which can be used to represent the speech phenomena we have to deal with and characterize the operations performed by the experimental equipment.

Consider the sample space containing all possible vector representations of the 18-channel quantized vocoder output. There are some 10^{16} vectors in this space, most of which of course are never used by an speaker in any context. Let us consider the sequence of patterns actually used by a given speaker over any short interval. Suppose a new sequence of patterns were artificially generated, in which some patterns differed from their original counterparts. It is reasonable to suppose that the quality of the resulting artificial speech would depend on:

- (i) How many patterns were changed relative to the total number spoken,
- (ii) How much each pattern was changed,
- (iii) Which patterns were changed,
- (iv) What text was being spoken,
- (v) Who was speaking and the time at which he spoke.

Let us refer to the combination of text, speaker and time of speaking as the input state. There is good evidence that for a given input state some patterns at least can be changed without an intolerable sacrifice of quality, but that if too many are changed by too much, serious degradation results.

This encourages us to take the view that the vectors corresponding to raw speech may be grouped into more or less well defined subsets in sample space, with the property that each vector in a raw speech sequence may be exchanged for any other vector lying in the same subset without intolerable degradation of the synthesized speech. We call these subsets linguistic exchange subsets.

This view ignores the possible effect of the context within which a given exchange occurs; i. e., the quality change brought about by exchanges within or between the subsets of a given configuration might depend not merely on the number of each type that occurs, but also on the order in which they occur. A more elaborate model could take this into account by considering subsets of allowable sequences of exchanges rather than subsets of allowable exchanges. We shall, however, adopt the simpler model to begin with, in the hope that the additional complexity will be found unnecessary.

It is hardly necessary to mention that use of the term "speech quality threshold" does not imply that this quantity is easy to define or measure, or indeed that it is a quantity at all. In practice its determination would involve the pooled value judgments of several skilled listeners.

We might add that it is possible, although not necessary, to postulate that each of the exchange subsets we have described has a linguistic correlate, something like a subphoneme, the concept being that there exists an alphabet of these subphonemes (perhaps 1000 or more in number) out of which phonemes and spoken words are constructed, just as written words are spelled out with letters. These subphonemes are abstract entities whose concrete realizations are the spectrum patterns obtained in 20 ms of vocoder output. Every vector pattern in an exchange subset has the same linguistic correlate, and these linguistic correlates or subphonemes are the building blocks of speech in the sense that they cannot be confused too much in analyzing the synthesizing speech without intolerable degradation of speech quality.

So far we have done little more than conjecture that a configuration of linguistic exchange subsets exists for any input state. How many subsets there are, to what extent they are localized in non-overlapping closed regions of sample space and how they shift as the input state changes are all questions that are clearly relevant to the design and feasibility of the speech compression system under study. They are furthermore questions that can be answered only by experimental study, and thus are sufficient to establish objectives for part of the experimental program.

Let us now turn to the sample space characterization of the operations performed by the analyzer-synthesizer, in an attempt to see what possibilities should be explored for optimally matching the machine to raw speech.

Regardless of how the reference library is selected, and whether it is held fixed or varied in an adaptive manner, the type of operation performed by the system on a particular input given a particular library is specified: The absolute distance between the input and each library reference (sum of channel differences without regard to sign) is calculated and the nearest reference is selected for transmission. Let us imagine that a certain number N of the points in sample space constitute the reference library during an interval when incoming raw speech patterns are to be classified. Let us divide the sample space into N distinct reference subsets, each containing the points closer to one of the reference points than to any of the other, distances being measured with the absolute yardstick mentioned above. These reference subsets are not hyperspheres but odd-shaped regions whose boundaries contain points equidistant from two or more references. The classification rule used by the machine in effect synthesizes speech by replacing each spectrum pattern of the original raw speech by the reference pattern corresponding to the subset into which the raw speech pattern falls. Thus, these reference subsets are also exchange subsets in the sense that the machine exchanges any pattern falling into one of the subsets for the subset reference pattern.

We may look upon the machine then as superimposing its configuration of reference subsets upon the configuration of linguistic subsets determined by the input state. If the reference subset configuration could be designed so that each reference subset covered one and only one linguistic subset at all times, we should expect optimum system operation. But this would be possible only if the linguistic subsets were separable by boundaries of the type set up by the machine, which is by no means known to be the case. Suppose, for example, that linguistic similarity depends on vectors being close in the sense that each coordinate (channel amplitude) should agree within one quantization level of the corresponding coordinate of some reference vector \underline{v}_o , and that not more than 5 coordinates or channels can show even this much variation. If the machine attempted to cover this linguistic subset by placing a single reference at \underline{v}_o , the nearest boundary would have to be at an absolute distance of 5 to include all of the points in the subset. But about half of the points within an absolute distance of 5 from \underline{v}_o would not belong to the linguistic subset because they differ from \underline{v}_o by more than one amplitude level in one or two channels. In other words when the reference subset is made large enough to include all points in the linguistic subset, many other points must be included as well, and these points of course might belong to other linguistic subsets. Since these exceptional points are well scattered throughout the reference subset, it is not clear how they can be separated easily from the others.

Just how serious this situation is remains to be determined. In any event, we can expect this type of machine to work well with a reasonable number of library references only if absolute distance is a good measure of linguistic similarity out to distances of say 4 or 5. Another objective of the experimental program, accordingly, is to find out whether this is the case.

The nature of the linguistic subset configuration determines the type of operation the system should perform for best results, and the quality of result that can be achieved. As mentioned above, the general model we have in mind is one in which the linguistic subsets are fairly distinct for a given input state (speaker, text, time of speaking), but shift around as the input state changes in such a way that the subsets always remain distinct although shapes and spacings may change.

For convenience we shall refer to parameters that characterize the linguistic configuration corresponding to a fixed input state as instantaneous parameters. Parameters that characterize the changes occurring as the input state changes may similarly be termed dynamic parameters.

The instantaneous subset configuration may be characterized grossly in terms of the average subset radius \bar{R}_L . This is the absolute radial distance from some central point within which most of the points of the subset may be found, on the average. As discussed above many other points which do not belong to a particular linguistic subset may lie within \bar{R}_L of its central point, and some of these may belong to other linguistic subsets. Thus, although two adjacent linguistic subsets may be disjoint, in the sense that they contain no points that belong linguistically to both, nevertheless their centers may be separated absolutely by less than $2 \bar{R}_L$. In this case we shall say that the two subsets although linguistically disjoint overlap absolutely.

Of course the instantaneous linguistic subset configuration may contain subsets with widely varying radii, in which case something more about the distribution of values of R_L than its average value will be needed, even for gross characterization. A rough picture of the instantaneous configuration can be obtained, however, by determining the average \bar{R}_L , the average separation of subsets \bar{S}_L in terms of \bar{R}_L and the approximate number of subsets.

As far as dynamic parameters are concerned, the most important gross characteristics are the average magnitudes of the changes in subset radius and location which occur as speaker, text and time of speaking are changed. These also may conveniently be expressed in terms of \bar{R}_L .

A quantity of considerable significance for system design is the time interval over which a given input state may be assumed to persist. It is quite possible that as a speaker proceeds through a long passage of homogeneous material, his linguistic subset configuration progressively changes so that subsets from later configurations may fall between subsets from earlier ones. Thus, if the input state is assumed to persist over too long an interval, a badly blurred configuration could result rather than the relatively distinct one we have associated with a given input state and have the right to expect on the basis of intelligibility. That is, the duration of a given input state is actually the time interval over which the configuration corresponding to a given speaker and given type of text remains fairly well separated. Since the configuration probably changes slowly with time, and varies somewhat with speaker and text; it may not be possible to specify this time interval precisely. On the other hand, even a rough average would be helpful in assessing the feasibility of an adaptive system which attempts to follow the changing configuration with a changing reference library.

Another conceivable difficulty here is that the duration of a given state is so short that not enough samples can be obtained to define the instantaneous configuration completely before it changes. In this event, experiments with short passages of text carefully selected for rapid definition of particular parts of the configuration may be necessary.

2.2 System Implications

Before proceeding to the discussion of experimental techniques for determining the parameters mentioned above, let us consider for a moment the system possibilities as they appear in view of the various situations that may exist in sample space.

The simplest situation is shown schematically in Figure 2.1a. Here the linguistic subsets are well separated for a given input state ($\bar{S}_L \gg 2\bar{R}_L$) and displacements are assumed small ($\bar{D}_L < \bar{R}_L$) as the input state changes. In this case a set of fixed references, one in each linguistic subset, will serve to classify input vectors correctly.

Figure 2.1b shows another situation in which the linguistic subsets, although stable, are closer together and irregularly shaped, so that for some subsets two references are needed for machine separation.

When the linguistic subsets shift by an amount comparable to \bar{R}_L as the input state changes, two further cases arise. In Figure 2.2a the shifts are larger than the mean subset radius ($\bar{D}_L > \bar{R}_L$) but smaller than half the mean distance between subsets ($\bar{D}_L < \bar{S}_L/2$), so that the possible positions of a given subset define a region which does not overlap the corresponding locus of positions of any other subset. Here the regions containing the loci can be separated by a machine using one reference per locus, but this may not be sufficient for good performance because a subset can shift so that it no longer contains its reference (as for reference r_i and subset L_{ij} in Figure 2.2a. In this case, although all vectors within a subset will always be replaced by a single reference vector, that reference vector may be too far from any of the subset vectors to bear a sufficiently close linguistic resemblance to them. (Of course, the shapes of the subsets may change more than is suggested in Figure 2.2a as the input condition changes.)

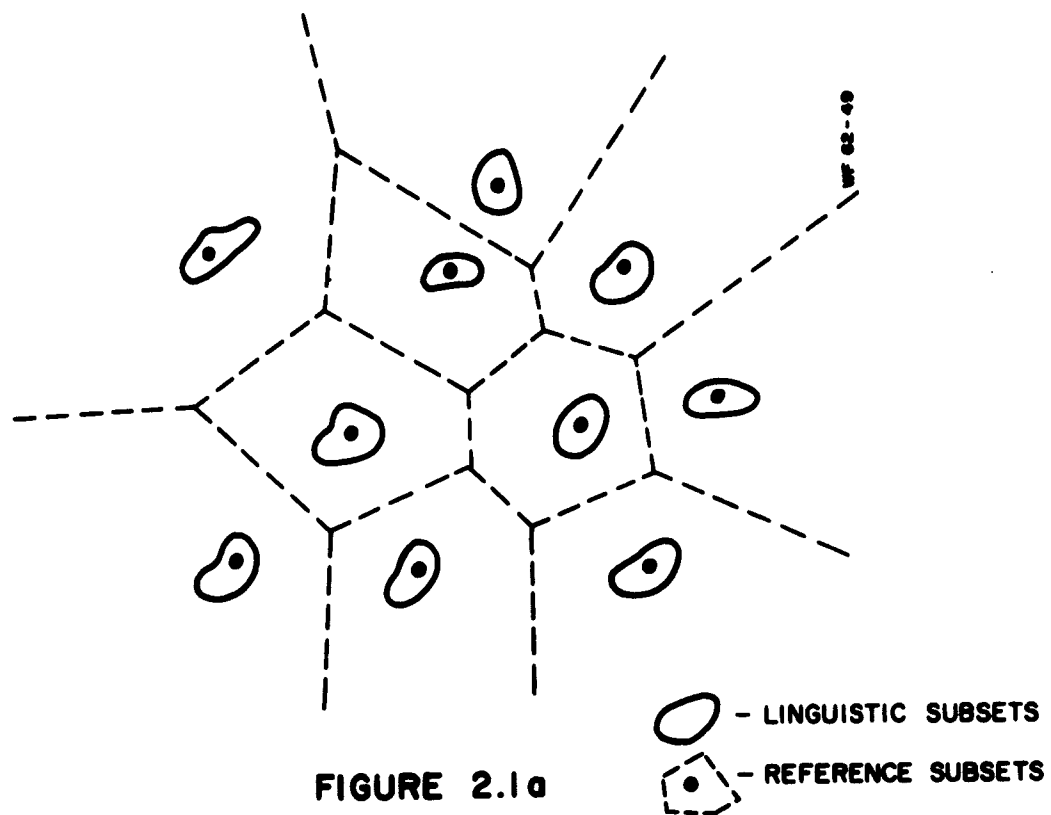


FIGURE 2.1a

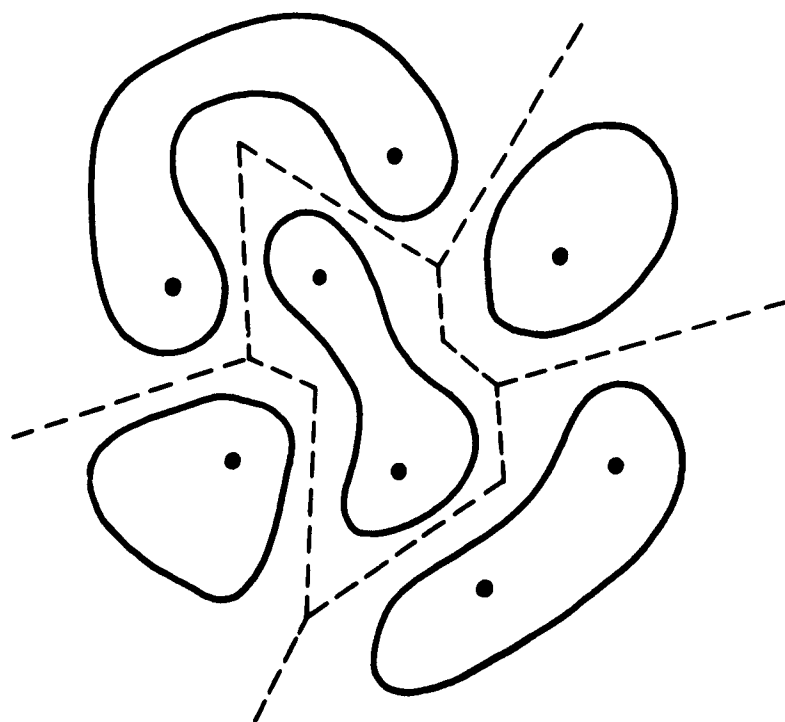


FIGURE 2.1b

Linguistic Subsets vs. Machine Subsets

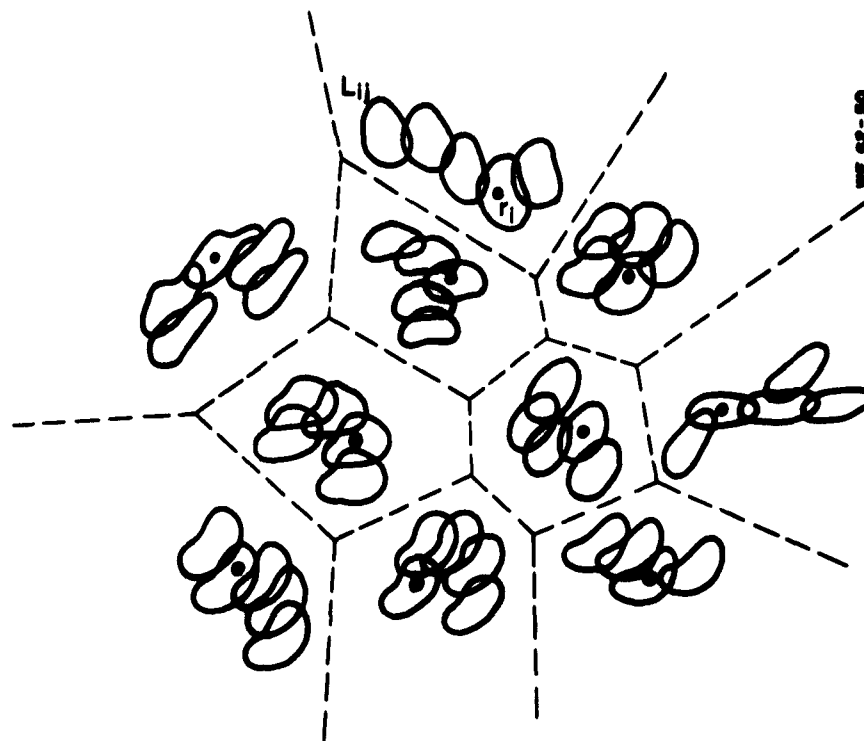


FIGURE 2.2a

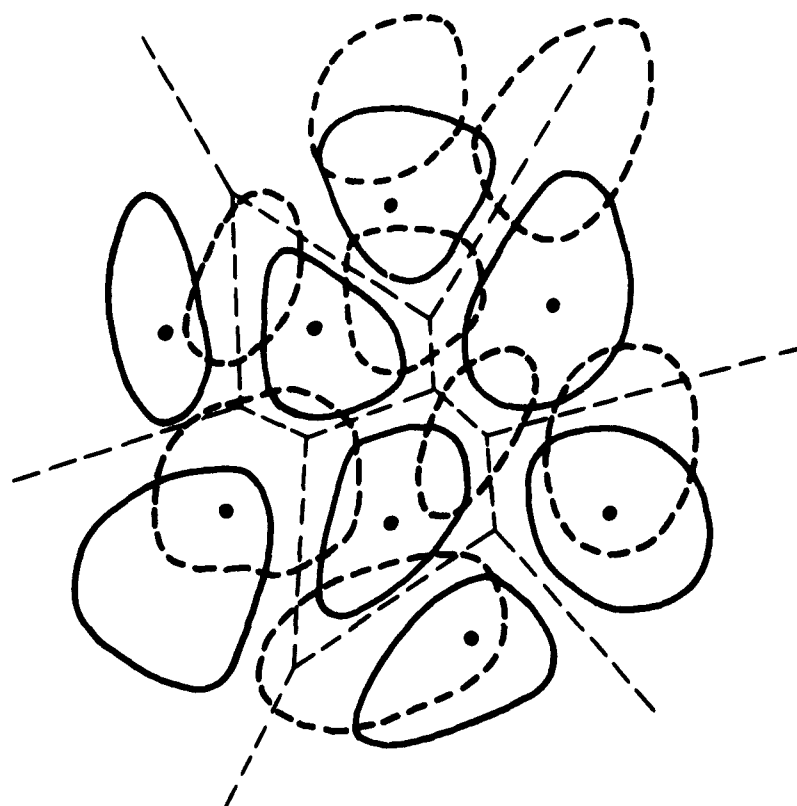


FIGURE 2.2b

Linguistic Subsets vs. Machine Subsets

The following system possibilities exist for the situation of Figure 2.2a:

(a) The "linguistic tolerances" may be sufficiently wide that one fixed reference per subset locus will serve as shown.

(b) If the linguistic tolerance is close, so that exchanges may be made only within subsets for any given condition, an adaptive system which derives its references from the subset pattern in use at the time would handle the situation.

(c) For a close linguistic tolerance a fixed grid of references, sufficiently dense that the tolerance requirement is met no matter what the subset displacements are, might be used.

In Figure 2.2b the subsets are close together ($\bar{S}_L \sim \bar{R}_L$) and the displacements are large ($D_L \sim \bar{R}_L$) so that subsets for one input condition overlap those corresponding to another input condition. Here two system possibilities are the dense fixed grid and the adaptive reference system mentioned above. A fixed reference system with one reference per subset would not work here, since if proper separation is achieved for one configuration, split subsets are bound to result when the same references are used for another input condition, as shown in the Figure.

2.3 Comparison of Reference Libraries

The AFCRL Voice Data Processing System is presently arranged to generate a library of references from raw speech input as follows:

- 1) A threshold $T \geq 0$ is set.
- 2) The individual channel quantized amplitude differences between the incoming pattern and each of the stored reference patterns are measured and totaled without regard for sign.
- 3) The incoming pattern is replaced by the nearest reference provided
 - a) No channel difference exceeds 1, and
 - b) The number of channels with a difference of 1 does not exceed T .
- 4) If both the above conditions are not met, the incoming pattern becomes a new stored reference.
- 5) In case of ties, the incoming pattern is replaced by the most frequently occurring of the tied references.

This generation scheme will produce a different set of library references each time a new input is applied. In order to make use of this feature for the studies we have in mind, it is convenient to introduce a measure of the distance between, or similarity of, two libraries generated from two inputs. Let the two sets of stored references be $\{x_i\}$ and $\{y_i\}$, with $i = 1, 2, \dots, N$ where N is chosen large enough to include all of the most frequently used references in both sets. Here we order the $\{x_i\}$ according to their frequency of occurrence, with x_1 the most frequent. The $\{y_i\}$ are then ordered with respect to the $\{x_i\}$ by starting with x_1 , choosing as y_1 the y that is closest

to x_1 , choosing as y_2 the y closest to x_2 , etc., i. e., so that

$$y_i = \min_j |x_i - y_j|, \quad i = 1, 2, \dots, N \text{ (in order)}. \quad (2.1)$$

where the bars denote absolute value. We are interested in the statistics of the quantity

$$d(x_i, y_i) \equiv d_i = |x_i - y_i|, \quad (2.2)$$

which measures the absolute distance between corresponding members of the two libraries. The average value of d_i may be taken as a simple measure of the similarity of the two libraries

$$d(X, Y) \equiv \bar{d} = \frac{1}{N} \sum_{i=1}^N d_i, \quad (2.3)$$

while its variance

$$\sigma_d^2 = \frac{1}{N} \sum_{i=1}^N (d_i - \bar{d})^2 \quad (2.4)$$

measures the uniformity of the fit.

This measure is actually based on a compromise between what one would like ideally and what can be computed easily. Ideally, to find the closest fit of the y -library to the x -library all (x, y) pairs should be searched for the closest pair, the next closest pair, etc., until all pairwise associations have been found. The above procedure does not do this; e. g., the y_i found in (2.1) might be closer to another x than to x_i . The pairwise associations produced by (2.1) depend on the order in which the x_i are taken, so that to avoid ambiguity a definite ordering of the x_i must be specified

Channel

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Number of Occurrences
2												2	3	1	1	1			348
3												1	1				1	3	301
4	1	1						2	1				2	1					202
5				1							2	2	2	1					195
6											1		1			1	3	2	182
7	3	1											3	1					178
8								1	1		1	2	1						167
9													1						162
10				2	1	1	2	2											157
11								1							1	3	2	2	145
12													1		2	1	2	1	133
13				1						2	1	1	2	1	1				125
14			1	2				2	2				1	1					116
15				2	1	2	2						1						102
16	1	2					1	2					2		1				86
17	1	3				1	2						2						84
18											1	1		1	2	2	2	2	81
19		1	1				2	3	2										74
20				1	1		1	2	2				1		1				74
21			2	2	1	2	2						2						72

Figure 2.3 Twenty Most Frequently Occurring References
 $T_M = 5$ (CPS - 29 August 1961)

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
2	0	10	10	6	12	8	8	7	16	13	9	7	13	14	11	13	11	17	13	15
3	0		12	10	6	12	8	5	14	9	7	11	13	12	13	13	9	15	13	15
4			0	10	13	6	8	7	16	15	13	9	7	14	9	9	17	9	11	15
5				0	12	10	6	7	14	17	13	5	11	12	13	13	13	17	13	13
6					0	14	10	7	13	7	5	13	15	14	15	15	7	17	15	17
7						0	12	7	16	17	13	11	13	14	9	9	17	15	15	15
8							0	5	14	13	11	7	9	12	13	13	13	11	11	15
9								0	9	10	6	8	8	7	8	8	12	10	8	10
10									0	17	15	15	13	4	11	11	19	13	7	7
11										0	6	15	16	17	16	18	6	16	14	20
12											0	12	14	13	12	14	6	16	12	16
13												0	8	13	12	14	12	14	12	14
14													0	11	16	16	18	8	10	12
15														0	13	9	19	17	9	3
16															0	6	18	12	8	14
17																0	20	16	14	10
18																	0	20	18	22
19																		0	10	18
20																			0	12
21																				0

Figure 2.4 Absolute Distances between References

When the two libraries are close in the sense that the average distance separating closest pairs is smaller than the average distance between members of the same library, the ideal and approximate ordering methods may be expected to give closely similar results. As an illustration we may use the data of Figure 2.3 with the ten even-numbered references taken as the x library and the ten odd-numbered ones as the y library. Both of these (artificial) libraries are already ordered according to frequency of occurrence. Figure 2.4 shows the absolute distances between all pairs. The following table shows the pairings produced by the two methods.

Pairs Given by (2.1)	d	No. of Occurrences	Closest Pairs	d	No. of Occurrences
2-5	6	543	2-5	6	543
4-7	6	380	4-7	6	380
6-3	6	483	6-3	6	483
8-9	5	329	8-9	5	329
10-15	5	259	10-15	4	259
12-11	6	278	12-11	6	278
14-13	8	241	14-13	8	241
16-17	6	170	16-17	6	170
18-19	20	155	18-21	22	153
20-21	12	146	20-19	10	148
$\bar{d} = 7.9 \quad \bar{d}_w = 6.9$			$\bar{d} = 7.9 \quad \bar{d}_w = 6.9$		

The average distance between members of the same library in this illustration is 12.3, which is considerably larger than the average separation of the pairs, in this case 7.9. As expected, therefore, the pairings are nearly identical; the only differences occur in the group 18 through 21.

A refinement of the measure of similarity introduced above may be obtained by using a weighted average in place of (2.3), with the weights dependent on the frequencies of occurrence of the pairwise associated patterns in the two libraries. Thus

$$\begin{aligned}\overline{d}_W &= \frac{1}{W} \sum_{i=1}^N W_i d_i, \\ W &= \sum_{i=1}^N E_i.\end{aligned}\tag{2.5}$$

When the W_i 's are taken as the numbers of occurrences shown in the table above, the weighted distance between libraries becomes 6.9.

Finally, we note that both d and \overline{d} satisfy the triangle inequality, for let

$$a_i = x_i - y_i \text{ and } b_i = y_i - z_i.\tag{2.6}$$

Then

$$d(x_i, z_i) \leq d(x_i, y_i) + d(y_i, z_i)\tag{2.7}$$

$$d(X, Z) \leq d(X, Y) + d(Y, Z)$$

become respectively

$$|a_i + b_i| \leq |a_i| + |b_i|\tag{2.8}$$

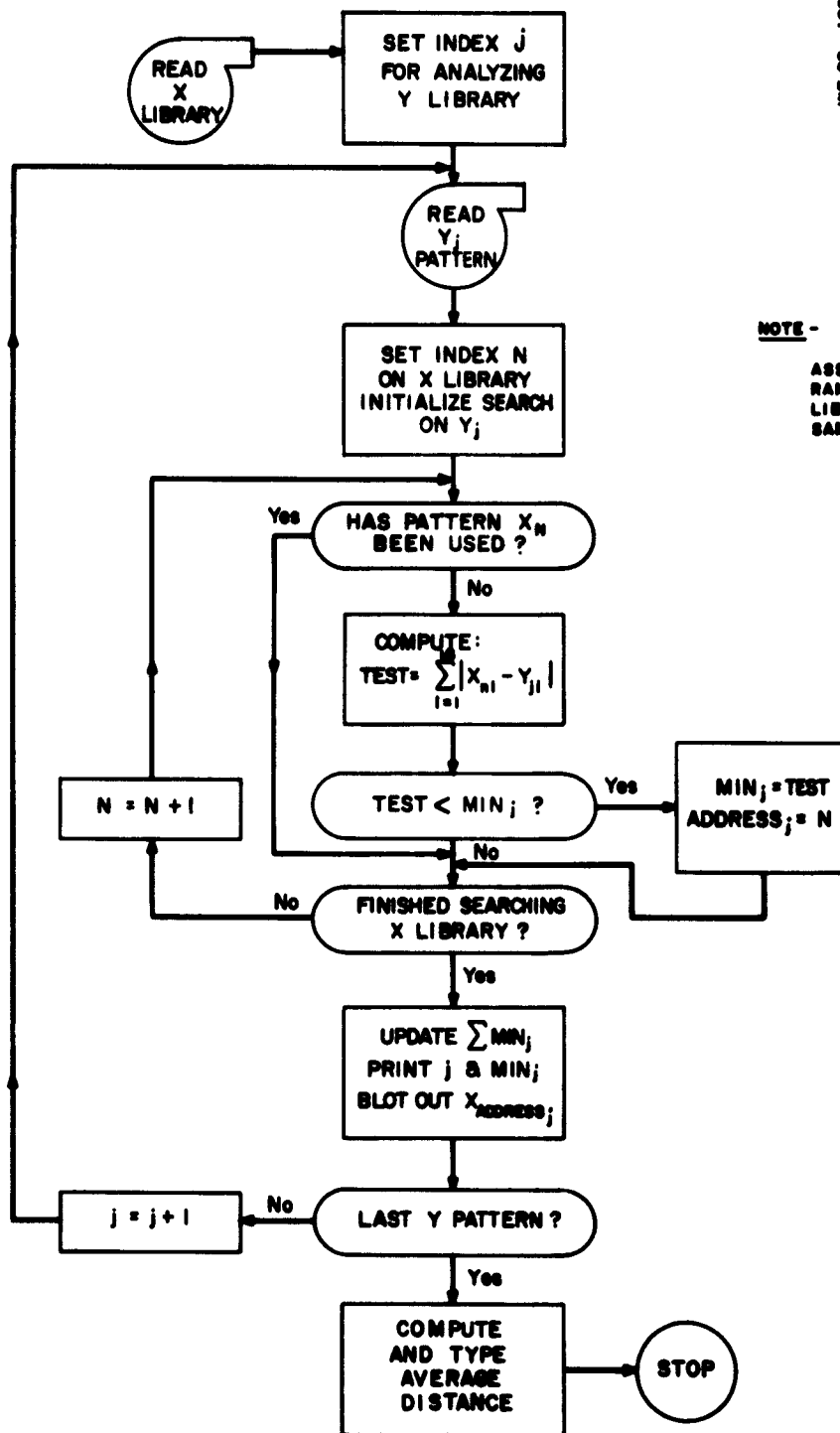
$$\sum |a_i + b_i| \leq \sum |a_i| + \sum |b_i|.$$

which are obviously true. Thus, for example, if two libraries are equidistant from a third, they themselves may be separated by not more than twice that distance, etc.

Flow Chart 2.1 shows how the distance between libraries may be programmed for machine computation. The program compares a rank ordered library X with another library Y which is rank ordered with respect to X by associating with each x_i taken in order the closest y . Once a vector y_i has been associated with an x it is deleted from the Y library. While making the comparisons the program computes the distances between the paired vectors, and after all comparisons prints out the average minimum distance. The paired vectors themselves are also available for printout.

On the Flow Chart $y'(n)$ contains the y -vector being paired at the moment, and $DIS(n)$ the minimum distance from the vector $y'(n)$ to its counterpart $x(n)$ in the X library.

Computer running time on a large scale digital computer (IBM 7090) should be in the vicinity of 2 minutes for 1000 vectors.



NOTE -

ASSUME X LIBRARY IS RANK ORDERED and X and Y LIBRARIES CONTAIN THE SAME NUMBER OF PATTERNS

FLOW CHART 2.1 Computation of Distance between Libraries

3. DETERMINATION OF EQUIPMENT ERRORS

3.1 Description of the Voice Data Processing Equipment

The AFCRL Speech Analyzer consists of an 18-channel vocoder and a multiplex⁽⁴⁾ which quantizes the amplitudes of each of the 18 bands of frequencies into 3 bits. This results in a 54 bit description of the instantaneous speech spectrum. In addition, the voice amplitude is quantized into 3 bits and the pitch frequency into 6 bits. An additional bit is reserved for indicating whether the sound is voiced or unvoiced. A 64 bit computer word thus describes the speech sound at an instant of time and 50 such samples are taken each second. Amplitude normalization is achieved by quantizing ratios with respect to the sum of the individual channel amplitudes, or area under the pattern.⁽⁵⁾

The voice data processor can compile a reference library of 54 bit spectrum patterns which it stores on a drum for future comparisons with the sequence of instantaneous spectra comprising speech. The method of comparison consists of computing the sum of the magnitudes of the channel amplitude differences between the input speech and the stored pattern, and matching the input speech pattern to the closest stored pattern.

The analyzer has basically two modes of operation. In the first mode it constructs a reference library of patterns from speech inputs, while in the second mode it transmits the memory address of the spectrum pattern which best matches the present input pattern. The match criterion is the computation described above.

In constructing the reference library of patterns, the machine compares the input spectrum with each of the stored patterns to establish the minimum error between the input and the best matching stored pattern.⁽⁶⁾ In the process of generating new references, the following criterion (if not met) results in the inclusion of the present spectrum in the reference library.

(a) In any of the 18 vocoder channels there must be at most one quantum level difference between the input spectrum and the stored reference with which it is compared.

(b) The sum of the magnitude of differences between input spectrum channel amplitude and stored reference channel amplitude must be less than a threshold T_M .

The threshold T_M can assume values from 0 to 7. By setting T_M equal to 0 the raw speech input can be accumulated and stored on the drum as if it were a sequence of references. Available outputs from the analyzer consists of:

(a) 21 channels of raw speech quantized as follows: 3 bits for each of the 18 vocoder channels, 1 bit for the voiced-unvoiced decision, 6 bits for pitch frequency, and 3 bits for voice amplitude.

(b) Speech in terms of the reference pattern stored on the drum.

(c) Speech in terms of the sequence of "closest" stored spectrum pattern numbers.

(d) Sequence of error numbers (the numerical value of the summed magnitude of channel differences).

(e) The number of occurrences of each of the stored references in the input speech segment.

(f) Rank ordering of the stored references.

Additional capabilities of the AFCRL facility include the capability of stretching speech to as much as 12 times its original duration by reading out (to the synthesizer) the sequence of stored patterns while repeating each pattern up to 12 times.

The readout from the drum can be connected to a vocoder synthesizer which will accept a 64 bit computer word, 54 bits of which represent the instantaneous spectrum as read from the drum. The synthesizer converts 64 bits to

20 analog signals and the binary voiced-unvoiced indication. The subset of 18 analog signals representing the spectrum channel amplitudes controls the gains of amplifiers which regulate the amount of energy transmitted at each instant as a function of frequency. The source of excitation of the set of filters used in the synthesis of speech is alternately the pitch frequency or the hiss generator. Switching between the two modes of excitation is governed by the voiced-unvoiced switch.

Thus, the AFCRL speech analysis equipment is a versatile research tool which not only permits a specific speech compression system to be tested and refined, but also facilitates basic research in many peripheral areas of speech analysis and perception.

3.2 Characterization of Sampling Errors

If the same input is repeatedly applied to the Voice Data Processing System (by means of a tape recording, for example) when operating with zero threshold, a new sequence of speech patterns may be produced at each trial. This results from the fact that the sampling at the vocoder channel outputs is controlled by a free-running clock, so that the sampling epoch, or point on the speech waveform at which the first sample is taken, may vary over one sampling interval. The sampling interval presently is 20 ms. The envelope detectors at the vocoder channel outputs have time constants of the order of 40 ms, so that a 3 db change of amplitude (one quantization level) can occur in less than 20 ms. In other words, as the sampling epoch shifts, as it is likely to do from trial to trial since it is uncontrolled, many patterns may change and a new and different set of library references may be generated.

Now the sampling interval and channel output time constants of the VDPS have been chosen to give an adequate representation of speech as determined from experiment. Presumably this is true regardless of sampling epoch. The situation here is analogous to sampling the amplitude of a time waveform, in which the one-dimensional amplitude samples of the latter correspond to the 18-dimensional samples of our vocoder output. In amplitude sampling we know that if the samples are taken frequently enough, then a good replica of the original waveform results when the sample values are processed in an appropriate low-pass filter. The result is not dependent on the sampling epoch or a particular sequence of amplitude values. Any sequence will serve as long as the rate of sample occurrence is not changed and the samples are taken consecutively in time. Analogously, although different pattern sequences may be obtained on different trials of the same input data with the VDPS, one sequence is as good as another as far as the effect on a listener is concerned.

In particular, the variations from trial to trial in the patterns obtained do not indicate the linguistic tolerances of the patterns. That is, for example, if the first non-zero pattern obtained on the first trial differs from the first non-zero pattern obtained on the second trial, this does not necessarily mean that these two could be exchanged, all other patterns in the two sequences remaining the same, without some degradation. On the other hand, the sampling rate built into the machine is somewhat faster than that dictated solely by sampling considerations, so that we should actually expect rather small trial-to-trial differences in corresponding patterns.

It would be very convenient if these trial-to-trial sampling variations were acoustically tolerable in the sense that the absolute distance between any two sets of patterns that can be generated by varying sampling epoch alone is much less than the linguistic tolerance threshold \bar{R}_g . If this is the case, sampling variations may be regarded as small errors and can be neglected in studying variations with text and speaker. Here the absolute distance is obtained by taking the sum of the absolute channel differences between each pattern in one library and the nearest pattern in the other and summing over all patterns in the first library, as described above.

Let us denote by $\bar{\epsilon}_g$ the average absolute distance between sets of patterns generated by sampling epoch variation alone. This quantity can be measured by recording a fairly long sample of speech and playing it repeatedly into the vocoder-analyzer, preserving the set of patterns generated on each trial. The absolute distances between all pairs of pattern sets are then calculated and the average of these is $\bar{\epsilon}_g$. The speech sample used for this experiment should be long enough so that the variety of acoustic patterns possible of utterance by a typical speaker is well represented. We do not know how long a sample this should be, but it would seem that about 20 seconds (1000 patterns) of text chosen for its representativeness should be

adequate. The effect of sample length can be determined by processing the patterns in two 10 second batches and comparing the $\bar{\epsilon}_g$ obtained for one batch with that obtained for the other and for the complete 20 second sample.

The number of times the recording should be rerun depends on how many distinguishable patterns can occur within one 20 ms sampling interval. This is unknown too, but the number is probably not more than two or three. If there are three distinguishable subintervals, the probability that at least one sample will be taken from each in N trials is given by

$$P_3 = 1 - (2^N - 1) 3^{1-N} \quad (3.1)$$

According to this, 9 trials will give a probability of 92 % that each of the three subintervals has been sampled at least once. For two distinguishable subintervals the corresponding probability is

$$P_2 = 1 - 2^{1-N} \quad (3.2)$$

3.3 Minimization of Synthesizer Errors

Once the analyzer sampling error $\bar{\epsilon}_s$ has been measured or estimated we have a basis for evaluating the performance of the synthesizer. Let us suppose that a speech sample has been read into the analyzer with the threshold set at zero and that the spectrum patterns occurring every 20 ms have been placed in storage. If the synthesizer is now used to manufacture artificial speech from the sequence of spectrum patterns, and the synthesized speech is fed back through the analyzer, ideally the spectrum patterns then obtained should correspond identically with the original patterns in storage. Identical correspondence will not occur, however, because of the sampling effect just discussed and also because the synthesizer even when optimally adjusted cannot reproduce the original waveform exactly.

The absolute distance between the stored set of patterns and the set produced by the synthetic speech immediately suggests itself as a measure of how well the synthesizer is adjusted. On the average this distance will be larger than $\bar{\epsilon}_s$ and the synthesizer should be adjusted to minimize it.

A technique for adjusting the synthesizer channel gains may be based directly on minimizing this absolute distance. After the synthetic speech has been passed through the analyzer, its spectrum patterns are placed in one-to-one correspondence with the stored patterns of the speech before synthesis in the usual way by associating each pattern in one set with the nearest pattern of the other set, thus obtaining say N pairs of patterns. Each individual channel is then examined by observing the distribution of the N differences occurring for that channel in the N pairs of patterns just obtained. If sign is preserved, a histogram of the differences would ideally be centered near zero. If the synthesizer gain for that channel is high or low the distribution will be skewed to the right or left by an amount proportional to the gain error. Of course, if the channel gains are badly misadjusted, considerable cutting and trying will be required to arrive at the proper adjustment.

4. RATE OF REFERENCE GENERATION

4.1 Occupancy Theory Model

Suppose the VDPS in the analysis mode is used to generate library references with a machine threshold $T_M = 5$. We would expect that at the beginning of speech, when the library is empty, the number of new references generated per second would be relatively large and that as time goes on and more references are accumulated in memory the number of new references generated per second would decrease, finally leveling off and approaching zero.

It is of considerable importance to know how this buildup occurs, particularly how long an interval is required to accumulate say 90 percent of the total number of references used by a speaker, and roughly what the total number is. The situation is complicated by the likely possibility that the duration of the instantaneous subset configuration used by a speaker may be shorter than the library buildup time. That is, a speaker may in effect shift from one reference library to another as time goes on, with a "dwell time" in any one library which is too short to achieve full buildup of that library. Thus, the study of the rate of reference generation is intimately related to study of the stability with time of a speaker's subset configuration.

A useful viewpoint for this study is provided by the following model, which although it involves considerable idealization of the actual case, does suggest some interesting interpretations of results.

Suppose we consider the random distribution of N indistinguishable objects among M cells or compartments. Each object is equally likely to be placed in any one of the M cells, independently of the placement of all other objects. There is no restriction on the number of objects in any particular cell. We are interested in how the number of cells containing at least one object depends on the number of objects N and the number of cells M .

We have in mind here, of course, an analogy with the VDPS analysis process, with the N objects being the first N spectrum patterns derived at the rate of 50 per second by the VDPS from a speech segment, and the M cells being M reference subsets each containing all patterns differing from the subset reference by not more than one level in at most T_M coordinates. The assumptions of independence and equal probability in the occupancy theory model are pretty certainly not met in the speech analogy, so that the former is an idealization of the actual case.

For the object-cell problem it is known* that for M and N large the average number n of non-empty cells is well approximated by

$$n = M(1 - e^{-N/M}), \quad (3.3)$$

i. e., the average number of empty cells decreases exponentially as the number of trials or objects to be distributed increases. (The actual distribution of the number of empty cells is approximately Poisson.)

4.2 Illustrative Examples

Figure 4.1 shows some data taken from runs made with a machine similar to the VDPS, with $T_M = 5$. The library buildup characteristic is roughly exponential, a reasonably good fit to the gross trend being obtained with

$$n = 166 (1 - e^{0.1t}) \quad (3.4)$$

shown as the solid curve in the Figure. By comparing (3.4) with (3.3) we note that they would coincide if $M = 166$ and $N = 16.6t$. Therefore, we may say that for the first 30 seconds or so the process behaves as if independent spectrum samples taken at the rate of 16.6 per second were being distributed randomly among 166 reference subsets. It is interesting to note that while the actual sampling rate of the machine was 50 samples per second, the

*See Feller, An Introduction to Probability and its Applications, Wiley, 1950. See 4.5.

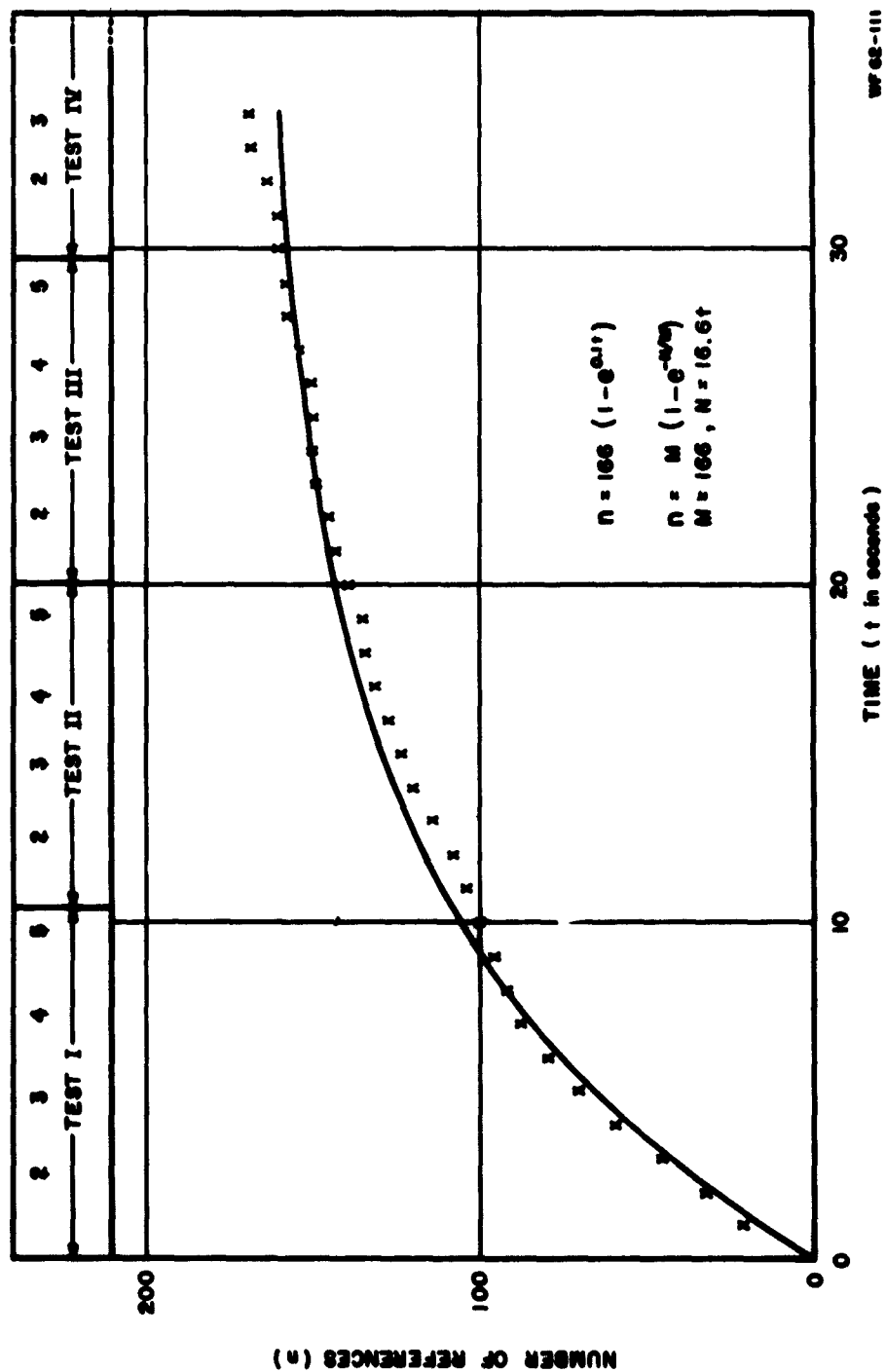


FIGURE 4.1 Rate of Reference Generation
 $T_M = 5$, 2-Bit Quantization
 (CPS 29 AUGUST 1961)

equivalent rate of independent sampling is only 16.6 per second or one-third the actual rate. Of course, the unrealistic assumptions behind the occupancy-theory model would prevent us from concluding with confidence that statistical dependence extends on the average only over three consecutive 20 ms speech samples, although that is what the analogy suggests. It also suggests that the library used by this speaker contains only about 166 reference spectra. If this were actually the case, and if speech synthesized using these references were of satisfactory quality, the implied transmission system could operate at a data rate of $50 \log_2 166 = 369$ bits per second, or perhaps even as low as $16.6 \log_2 166 = 123$ bits per second. The chain of inferences here is a long one with several weak links, however, so that at present these numbers must be regarded as highly speculative.

Actually when we look closely at the experimental data of Figure 3.8 we see that a sequence of exponential segments might fit the data better than the single segment, which may be evidence that the speaker periodically makes slight changes in the library he is using. The run from which the experimental points were taken consisted of a number of separate tests during each of which four 2-3 second sentences were spoken, followed by a fifth phrase "End of Test ____". The test intervals and individual sentence starting times are indicated on the Figure. We note that at the ends of Tests I, II, and III there are breaks in the trend of the data. Another break occurs at the beginning of the fourth sentence of Test III. The data continues its upward trend during Text IV and succeeding tests not shown here (some garbling occurred later).

Figure 4.2 shows another buildup characteristic, taken with a smaller machine threshold ($T_M = 3$) and extending over a longer time interval. We note that the characteristic is approximately equivalent to the buildup in the number of subsets occupied if patterns independently selected at a rate of 23.2 per second were being randomly distributed among 3100 reference subsets.

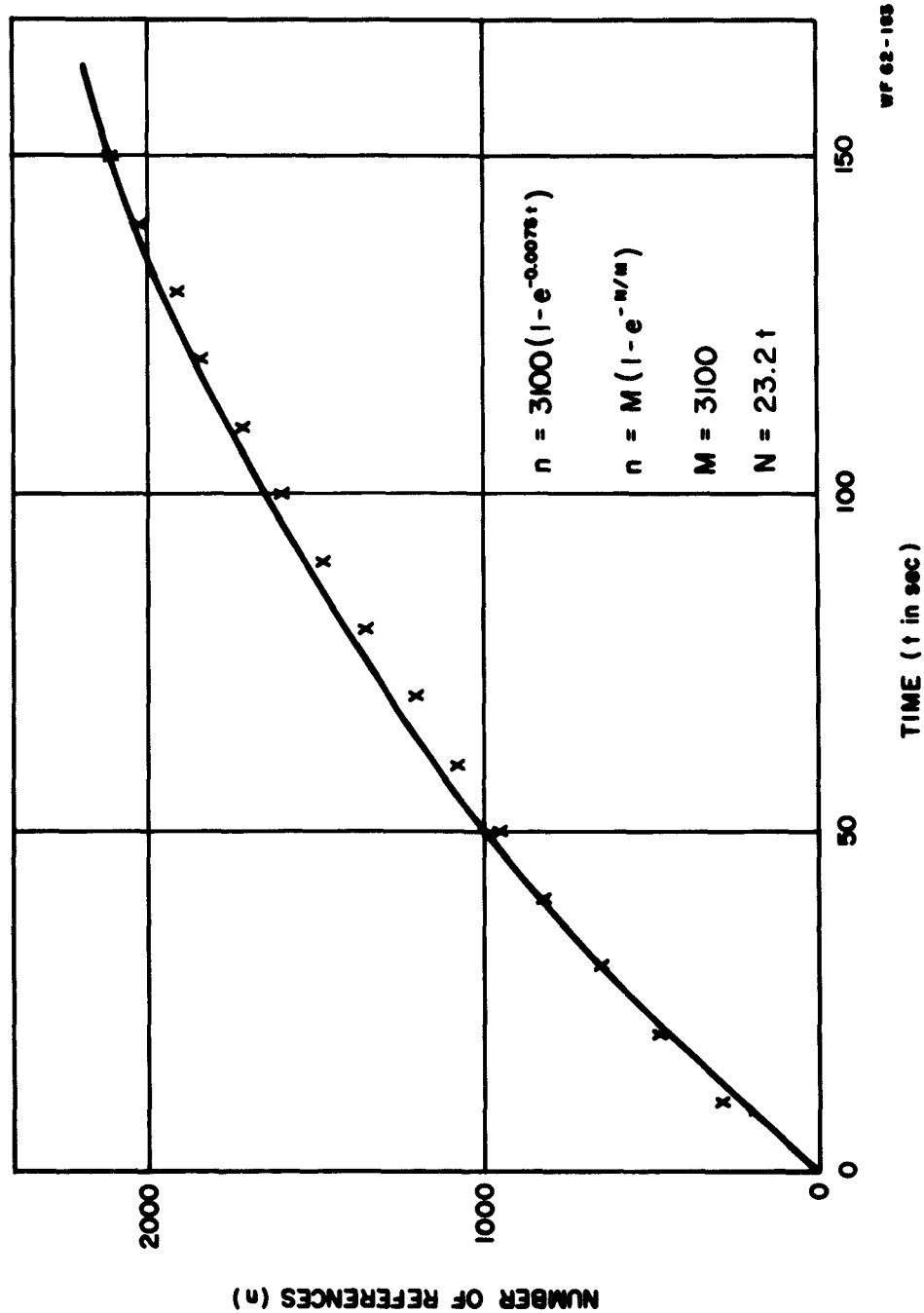


FIGURE 4.2. Rate of Reference Generation, $T_M = 3$, 2-Bit Quantization
 (R.J. McB., Selection 6, JULY 11, 1962)

Comparing this latter case with the former, we note that in one the patterns are being distributed among a few relatively large subsets ($T_M = 5$, $M = 166$), whereas in the other the distribution is over a larger number of smaller subsets ($T_M = 3$, $M = 3100$). It follows from the rule used by the analyzer to define subsets that subsets with $T_M = 5$ and $T_M = 3$ contain a total of about 330,000 and 7,200 points respectively, each point corresponding to a different pattern. Thus, the total number of points "spanned" by the array of subsets is $166 \times 330,000 = 55,000,000$ in one case and $3,100 \times 7,200 = 22,000,000$ in the other. Since these agree within an order of magnitude, we may conjecture that the total volume of 18-dimensional space utilized is roughly the same in the two cases. There are some $4^{18} = 10^{12}$ points altogether in the space, so that the percentage utilization is on the order on one-hundredth of one percent.

The equivalent rate of independent sampling in the second example with $T_M = 3$ turned out to be 23.2 per second, in contrast to 16.6 per second when $T_M = 5$. If we assume crudely that the reduced rates observed with $T_M = 3$ and 5 may be attributed to a tendency for successive vectors to remain in the same subset, we would expect the observed rates to depend inversely on the subset diameters. The ratio of subset radii for $T_M = 3$ and 5 is in the order of $(7,200/330,000)^{1/18} = 0.81$, while the corresponding ratio of rates is $(16.6/23.2) = 0.72$, indicating agreement to within about 10 percent. (Of course, the subsets in question here are defined in such a peculiar way that the definition of their radii is somewhat arbitrary.)

It is clear that further study and model-making is required for satisfactory interpretation of these buildup characteristics. They appear to contain valuable clues concerning the sample space configuration and its stability with time. The occupancy-theory model may be capable of modification to include the effects of periodically increasing the number of cells among which the trials are distributed, in which case its value as an interpretive aid would probably be enhanced.

In addition to the buildup of the number of machine references as illustrated here, it would be desirable to study buildup of the number of absolutely disjoint subsets, i. e., the number of new references obtained by clustering the machine references as discussed above in Section 5.

5. A LIBRARY REDUCTION TECHNIQUE

5.1 Spherical Clustering

Here we shall discuss a specific technique which operates on a set of vectors and produces subsets each containing vectors closer to each other than to those in other subsets; i. e., it is a clustering technique. With this technique we may examine the structuring of sample space produced by any particular input condition.

The starting point for this distribution study is the list of reference patterns stored by the VDPS when a speech sample has been fed in at the analyzer input, with the machine threshold set at a particular value T_M . Each raw speech pattern then differs by not more than one quantization level in not more than T_M channels from some one of the references. Note that if spheres of radius T_M are constructed about each reference as a center, many points of the space will be common to two or more spheres; i. e., the spheres are not absolutely disjoint. We may regard the references generated in this way as indications of where the raw speech patterns are concentrated in the space, but not necessarily as the optimum set of references for best separation of exchange subsets in view of the nearest-absolute-distance criterion used by the machine in making classification decisions.

When the spherical clustering technique to be described is applied to a set of stored references, the objective is to find out which references are close enough to each other that if each cluster is replaced by a reduced number (perhaps a single one) of suitable chosen new references, then the nearest-absolute-distance decision rule operating with the new references will produce raw speech pattern replacements which are within the linguistic tolerance. The assumption behind this is that linguistic tolerance depends on absolute distance, as it must if the VDPS is to work satisfactorily with a reduced number of fixed references.

The spherical clustering technique separates a set of vectors into subsets each of which contains vectors close to each other in the sense that the absolute distance between any pair of vectors in a given subset is less than some prescribed distance S . Let us call the set of vectors within a distance S of a given vector V the S -neighborhood of V and denote it $S(V)$. The procedure for finding the subsets required is quite simple and obvious. First, the vector V_1 having the largest S -neighborhood $S(V_1)$ is found. The vectors in $S(V_1)$ although all close to V_1 , may not be close to each other, so the next step is to find the vector $V_2 \in S(V_1)$ which itself has the largest S -neighborhood $S(V_2)$ within $S(V_1)$. The vectors in $S(V_2)$ are now all close to both V_1 and V_2 but still may not all be close to each other. The process is continued until the remaining vectors in $S(V_1)$ are exhausted. The result is a subset C_1 of vectors V_1, V_2, \dots each of which is within an absolute distance S of each of the others in C_1 .

To find the next cluster C_2 the entire process is repeated starting with the original set of vectors with those belonging to C_1 deleted. The remaining clusters are found in a similar way.

5.2 Illustrative Examples

As an illustration of the method, the 20 references shown in Figure 2.3 were clustered. These were the 20 most frequently used references generated by a machine of the type we are considering with a threshold setting of 5 when a speech sample of about 1 1/2 minutes duration was analyzed. Figure 2.4 shows the absolute distances between each pair of references, and Figure 5.1 is a histogram showing the distribution of these distances. A threshold distance S of 10 was arbitrarily chosen at first as the distance within which all vectors belonging to the same cluster must be spaced. Figure 5.2 shows the distances after thresholding. Inspection reveals that Reference 9 has the largest S -neighborhood, and that References 4, 2, 5, 8, and 13 constitute the sequence of selections remaining to complete the definition

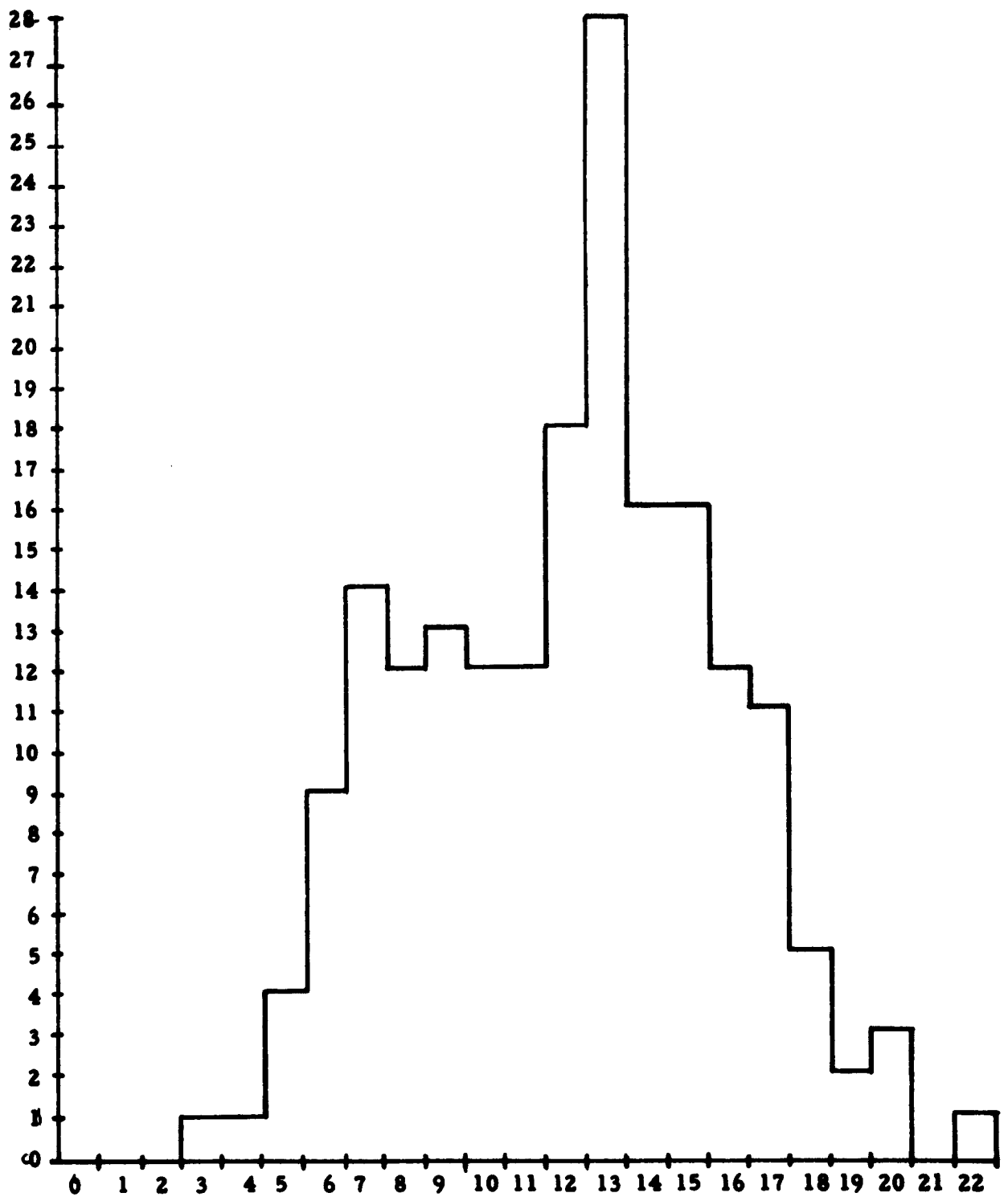


Figure 5.1 Histogram of Absolute Distances

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
2	0	0	0	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1
3		0	1	0	0	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1
4			0	0	1	0	0	0	1	1	1	0	0	1	0	0	1	0	1	1
5				0	1	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1
6					0	1	0	0	1	0	0	1	1	1	1	1	1	1	1	1
7						0	1	0	1	1	1	1	1	1	0	0	1	1	1	1
8							0	1	1	1	1	0	0	1	1	1	1	1	1	1
9								0	1	0	0	0	0	1	0	0	1	0	0	0
10									0	0	0	0	1	0	1	1	1	1	0	0
11										0	1	1	1	1	1	1	0	1	1	1
12											0	0	1	1	1	1	1	1	1	1
13												0	1	1	1	1	1	1	1	1
14													0	0	1	1	1	0	1	1
15														0	1	1	1	1	0	0
16															0	0	1	1	0	1
17																0	1	1	1	0
18																	0	1	1	1
19																		0	1	1
20																			0	1
21																				0

$$C_1 = (9, 4, 2, 5, 8, 13) \quad C_4 = (17, 7, 16)$$

$$C_2 = (3, 6, 11, 12, 18) \quad C_5 = (14, 19)$$

$$C_3 = (15, 10, 20) \quad C'_3 = (21)$$

Figure 5.2 Distances after Thresholding

$d \leq 10 \rightarrow 0$

$d > 10 \rightarrow 1$

of the first cluster C_1 . With these references eliminated, the procedure is repeated to find C_2 , and so on, the clusters obtained being as shown at the bottom of the Figure. In the derivation of these, the rule is used that if more than one reference has the same maximum S-neighborhood, the highest ranking one is chosen. The last cluster turns out to be a single vector (21) which has been labeled C_3' because of its proximity to C_3 , and has been included with C_3 in the calculations to follow.

In order to see how good these clusters are, we may choose a single representative vector for each and inspect the intra- and intercluster distances. Figure 5.3 shows the vectors arranged by clusters and the representative new references chosen for each. The rule used for the latter was to average each channel and round off to the nearest integer. In Figure 5.4a the distances from each of the new references to each of the old references are tabulated. We note that a machine using the nearest-absolute-distance rule with the new references as the stored library would classify all of the old references correctly (in the case of Reference 9 there is a tie between R_1 and R_2 which would be resolved in favor of cluster R_1 since C_1 has a larger population than C_2). We see from Figure 5.4b that the average cluster radius (4.5) is significantly smaller than the average distance between new references and adjacent clusters (13.0). Figure 5.4c shows the separations between all pairs of new references, averaging 12.6.

Of course these results are illustrative only and are strongly affected by the machine threshold T_M used to obtain the original 20 references and the threshold S used to define the clusters. In particular when these thresholds are large, as in the example, ($T_M = 5$, $S = 10$), many raw speech patterns which were originally associated with a given reference may turn out to be replaced by a different new reference than the new reference which replaces the given reference. Figure 5.5 makes this clear. Pattern X

		Channel																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
C_1	2												2	3	1	1	1		
	4	1	1							2	1			2	1				
	5				1							2	2	2	1				
	8								1	1	1	2	1						
	9												1						
	13				1						2	1	1	2	1	1			
Avg.										1	1	1	1	2	1				
		R_1																	
C_2	3												1	1				1	3
	6										1		1				1	3	2
	11								1							1	3	2	2
	12												1			2	1	2	1
	18										1	1			1	2	2	2	2
Avg.													1		1		1	2	2
		R_2																	
C_3	10				2	1	1	2	2										
	15				2	1	2	2						1					
	20				1	1		1	2	2				1		1			
	21			2	2	1	2	2						2					
Avg.				1	2	1	1	2	1	1				1					
		R_3																	
C_4	7	3	1											3	1				
	16	1	2					1	2					2		1			
	17	1	3					1	2					2					
	Avg	2	2					1	1					2					
		R_4																	
C_5	14			1	2					2	2			1	1				
	19		1	1						2	3	2							
	Avg		1	1	1					1	3	2		1	1				
		R_5																	

Figure 5.3 Choice of New References (S=10)

	C ₁					C ₂					C ₃					C ₄					C ₅				
	2	4	5	8	9	13	3	6	11	12	18	10	15	20	21	7	16	17			14	19			
R ₁	7	5	5	3	6	4	9	11	14	12	12	15	13	12	14	9	12	12			8	12			
R ₂	9	13	13	6	6	12	5	3	4	2	6	15	13	12	16	13	12	14			14	16			
R ₃	16	14	14	12	9	15	14	16	17	15	21	4	4	7	5	16	13	11			9	13			
R ₄	12	8	12	12	7	13	12	14	17	13	19	12	12	11	13	6	3	5			15	13			
R ₅	15	7	13	11	10	10	15	17	18	16	20	15	15	10	16	13	14	16			4	4			

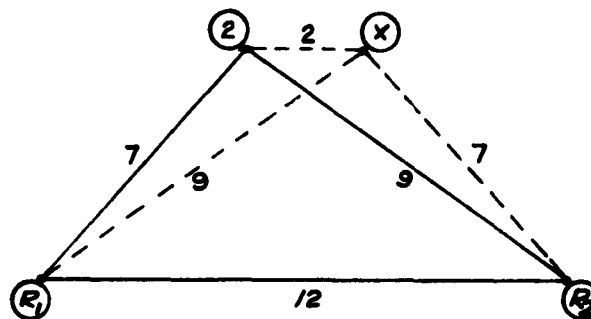
Figure 5.4a. Distances Between New References and Old References (S=10)

	R ₁	R ₂	R ₃	R ₄	R ₅
R ₁	0	12	13	11	10
R ₂		0	15	13	16
R ₃			0	12	11
R ₄				0	13
R ₅					0

Figure 5.4c Distances Between New References (S=10)

	C ₁	C ₂	C ₃	C ₄	C ₅
R ₁	5.0	11.6	13.3	11.0	10.0
R ₂	10.7	4.0	13.3	13.0	15.0
R ₃	13.3	16.6	5.0	13.3	11.0
R ₄	10.7	15.0	11.7	4.7	14.0
R ₅	11.0	17.2	13.3	14.3	4.0

Figure 5.4b Average Distances Between New References and Clusters (S=10)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
R_1									1	1	1	1	2	1				
R_2													1		1	1	2	2
2												2	3	1	1	1		
X												2	3	1	1	1	1	1

Figure 5.5 Vector X Is Close to 2 But Would be Classified Differently

was originally associated by the analyzer with reference pattern 2, from which it differs by not more than 1 per channel in not more than 5 channels. After clustering the reference patterns, pattern 2 becomes part of the first cluster represented by the new reference R_1 , and indeed it is closer to R_1 than to any other of the new references. Pattern X, on the other hand, is closer to R_2 than to R_1 . In the operation of the machine in the classification mode, therefore, the two patterns 2 and X which are within a distance 2 of each other are exchanged for two patterns R_1 and R_2 which are separated by a distance of 12. Of course, both 2 and X are replaced by new patterns only 7 units distant in each case which is comparable with the radius of the subsets generated by the clustering scheme. Thus, if it is linguistically tolerable to represent subsets of this radius by single references near their centers, the replacements mentioned may also be linguistically tolerable. The fact that this order of magnitude is probably not tolerable means that the original data should have been taken with a smaller machine threshold T_M and a smaller clustering threshold S.

In order to observe the effect of changing the clustering threshold S, the 20 vectors of Figure 2.3 were clustered again with a threshold of 8 rather than 10. The clusters produced at the two threshold values are compared in the following table:

	<u>S=8</u>	<u>S=10</u>
C_1	2-5-8-9-13	2-4-5-8-9-13
C_2	6-11-12-18	3-6-11-12-18
C_3	10-15-21	10-15-20
C_4	16-17	7-16-17
C_5	4-14-19	14-19
C_6	20	21
C_7	7	
C_8	3	

Referring to Figure 5.3, we note for example that the changes brought about in C_1 and C_5 , viz. transfer of reference 4 from C_1 to C_5 seems reasonable because reference 4 has a definite formant structure which is lacking in the remaining vectors of C_1 , and which better fits the structure pattern of references 14 and 19. Figures 5.6 and 5.7 show the choice of new references for $S=8$, the average distances between new references and clusters, and the distance between references. It would appear that $S=8$ is somewhat more satisfactory as a clustering threshold than $S=10$ in this case.

While these examples are illustrative only, if the technique had been applied to a realistic and complete set of vectors, whether raw speech patterns or machine-produced references, the next step would be to use the VDPS to resynthesize the original speech in terms of the reduced number of new references. The nature of the procedure is such that most of the original speech vectors will be close to one or another of the new vectors, so that the resulting resynthesized speech is nearly optimum for the reduced number of references. If it is not of adequate quality, the spherical clustering must be repeated with a smaller cluster diameter.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
C_1	2												2	3	1	1	1		
	5				1							2	2	2	1				
	8									1	1	1	2	1					
	9													1					
	13				1						2	1	1	2	1	1			
Avg											1	1	1	2	1				R_1
C_2	6											1		1			1	3	2
	11									1						1	3	2	2
	12												1			2	1	2	1
	18										1	1			1	2	2	2	2
Avg																1	2	2	R_2
C_3	10				2	1	1	2	2										
	15				2	1	2	2						1					
	21			2	2	1	2	2						2					
	Avg			1	2	1	2	2	1					1					R_3
C_4	4	1	1							2	1			2	1				
	14			1	2					2	2			1	1				
	19		1	1					2	3	2								
	Avg		1	1	1				1	2	2			1	1				R_4
C_5	16	1	2					1	2					2					
	17	1	3				1	2						2					
	Avg	1	2					1	1					2					R_5
C_6	3												1	1				1	3
C_7	7	3	1											3	1				
C_8	20				1	1		1	2	2				1		1			

Figure 5.6 Choice of New References ($S=8$)

	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	R ₈
C ₁	4.4	12.2	13.6	10.8	10.2	8.8	9.6	11.4
C ₂	11.8	3.8	17.8	16.8	14.8	7.8	14.5	14.8
C ₃	13.0	16.0	3.0	14.3	11.3	13.7	14.0	9.3
C ₄	9.3	15.7	14.0	4.7	11.0	13.3	11.3	10.3
C ₅	11.0	15.0	12.0	14.0	3.0	13.0	9.0	11.0
C ₆	8.0	7.0	14.0	14.0	11.0	0.0	12.0	13.0
C ₇	8.0	15.0	16.0	12.0	7.0	12.0	0.0	15.0
C ₈	13.0	14.0	9.0	9.0	10.0	13.0	15.0	0.0

Figure 5. 7a Average Distances Between New References
and Clusters (S=8)

	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	R ₈
R ₁	0	13	14	10	9	8	8	13
R ₂		0	17	17	14	7	15	14
R ₃			0	12	11	14	16	9
R ₄				0	11	14	12	9
R ₅					0	11	7	10
R ₆						0	12	13
R ₇							0	15
R ₈								0

Figure 5. 7b Distances Between New References (S=8)

5.3 Machine Computation

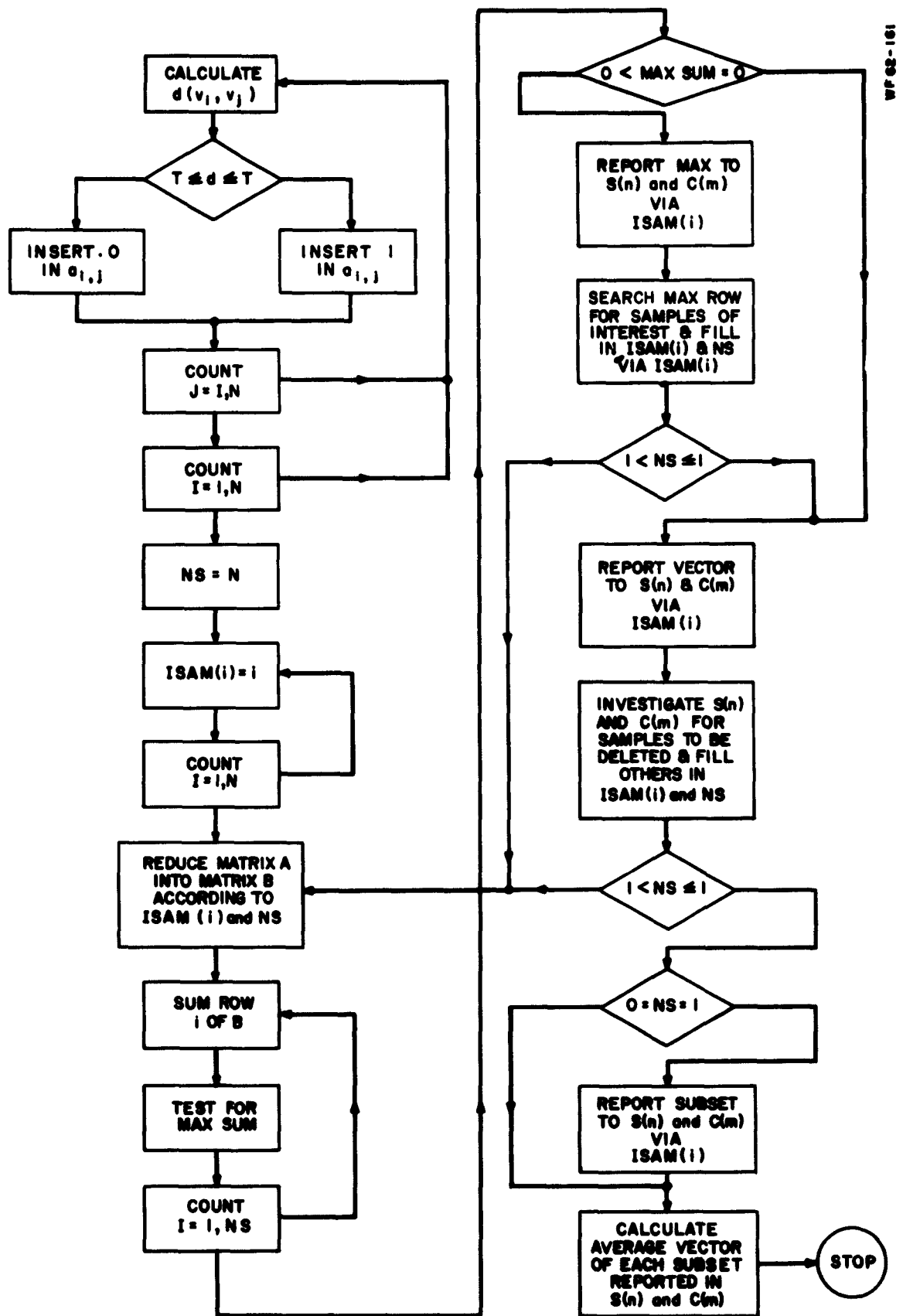
The spherical clustering technique described above was programmed for the RECOMP II Computer and tried on sample data from the VDPS.

Owing to computer memory limitations, the program was designed to handle a maximum of 245 reference patterns. Flow Chart 5.1 shows how the computation was programmed.

Data supplied by C. P. Smith in serial format were processed to eliminate repetitions, and the 245 most frequently occurring patterns were selected for clustering. Distances between all pairs of patterns were calculated and compared with a threshold of 10, selected by inspection of the data. The serial data were taken with 2-bit quantization and a pattern deviation of 3. The original text was as follows: "THIS IS R. J. MACBAIN SELECTION SIX. HERE IS THE SIXTH SELECTION. IT IS A COLLEGE LECTURE ON AN ASPECT OF LANGUAGE. WE TEND TO THINK OF A LANGUAGE AS AN ACCURATE STABLE THING WHICH..."

The clustering computation was carried to the point where 188 of the patterns were grouped into 36 clusters containing from 3 to 12 patterns each. Three of the clusters formed are shown in Figures 5.8a, b, c to illustrate the nature of the results. Well over 100 hours of RECOMP II time was consumed by even this modest program, indicating that a larger and faster computer would be required for a full scale run.

In a continuation of work with these data, another threshold might be tried on all or a part of the data to see if the clustering could be improved (although the present results seem to be rather good), after which a representative vector from each cluster would be used to make up a reduced-size library for resynthesis of the original speech, as discussed above.



FLOWCHART 5.1. Spherical Clustering Computation.

DATA: Cluster No. 1

[illegible]

DATT: Cluster No. 2

[illegible]

[illegible]

6. LIBRARY VARIABILITY

6.1 Analysis of Variance

The index of library similarity introduced in 2.3 makes it possible to study quantitatively the significance of variations in the sample space configuration produced by speaker and text differences. We need to do this in order to assess the feasibility of using a fixed reference system, as discussed above in 2.2. We need to know how much variability there is in the reference libraries produced under different input conditions, and whether the variability can reasonably be attributed to changes in the input condition or whether it is normally what would be expected in view of random sampling from fixed distributions.

In taking data for analysis using this measure of library similarity, one reference library is chosen as a standard against which all others obtained on different trials are compared. A single number representing the average distance between the library obtained and the standard results from each trial, and it is the variability of this quantity that is of interest. It is necessary to refer each library to a standard in order to evaluate the measure of similarity, which involves differences between vectors in one library and those in another. The choice of standard is arbitrary but should be governed by the peculiarities of the measure of similarity used. As noted above in 2.3, the measure used increasingly departs from an ideal measure of closeness as the average displacement between libraries increases relative to the average separation of vectors within the libraries. This implies that the standard should not be far away from the libraries to be studied. Since we want to exhibit the variations as clearly as possible, a good choice would seem to be one of the observed libraries itself. If the mean of all libraries observed were chosen, the variation would not show up as well because the

absolute distances involved in the measure of similarity would tend to be roughly the same even though the libraries were displaced considerably; e. g., two libraries on "opposite sides" of the standard might be exactly the same absolute distance away from the standard and yet be twice this distance apart from each other, so that the variation between them would be totally obscured.

We shall adopt as a general procedure then the use of the library obtained on a particular run as the standard against which others are compared for the purpose of analyzing their variability. The particular run chosen will usually be the first of the series to be studied, although special circumstances may dictate a different choice.

A vast literature exists on techniques for analyzing the variability of such experimental results, and for designing experiments to bring out certain types of variations more clearly. It would be neither possible nor appropriate to attempt here a summary of this material, usually found in standard statistical texts under the headings of Analysis of Variance and Design of Experiments.* Accordingly, we shall limit the discussion in this section to some general indications of the types of assumptions involved and results obtainable when these methods are used, and in the following sections suggest some more specific procedures for the particular problems of the present application.

The elementary techniques for analyzing variance, which we may expect to find the most useful for our application, are based on comparing the variations among different families of trial observations with the variations within families. If the variation among families is significantly larger than that within families, we may conclude that there is a "family effect" in the data; e. g., in our application that changing the parameter set does indeed produce a significant change in system performance.

*See for example Scheffe, The Analysis of Variance, Wiley 1959, and Kempthorne, The Design and Analysis of Experiments, Wiley 1952.

The techniques for making significance tests of this kind depend on the fact that if the trial observations are independent samples from a common underlying Gaussian distribution, then the variation between families is statistically independent of the variation within families. Under this condition the sums of the squared variations of these two types have independent chi-squared distributions, and their ratio has a Fisher-Snedecor F distribution, which is tabulated.

Thus it is easy to find the probability that any observed value of the ratio would be obtained purely by chance on the hypothesis that all data were drawn from a common underlying Gaussian distribution, so that all variations were due solely to sampling fluctuations. The level of significance of a value is the probability that a purely random sample would deviate even further from the mean of the F-distribution than does the value in question. However, if it turns out for example that the significance level of a particular ratio value were only 1 % on the assumption that sampling error was the only operative factor, we might have good reason to reject this hypothesis.

These elementary techniques depend for their validity on the assumption that in the absence of family effects the data would be homogeneous in the sense that the observations could be then regarded as samples from the same Gaussian population. This is often not a good assumption, however, and the whole basis of the analysis is threatened unless ways can be found to circumvent this difficulty. Fortunately under certain conditions the F test is rather insensitive to this assumption. Moreover, many techniques of experimental design are available which make it possible to use the same type of significance test even when the populations from which the samples are drawn are non-Gaussian and unknown.

To illustrate the underlying idea common to many of these "design of experiment" techniques, suppose we wish to determine whether there are

significant differences in the libraries used by a given speaker when he reads three different types of text. We may envision a number of trials, in each of which first a sample of one type of text is read, then a sample of the next type and finally a sample of the last type. This is repeated several times. Now in analyzing the results to determine whether there is significant variation between types of text, we first make the hypothesis that there is none (the null hypothesis), and then reason that if indeed there were none, the data on each trial should be homogeneous. However, if the types of test are always read by the speaker in the same order on each trial, it is clear, for example, that fatigue or training effects might influence the third type, which is always the last one read, more than the others. That is, there might well have been systematic variations in the data due to fatigue even if the choice of text type actually made no difference at all. The assumption of homogeneity and the test of significance would therefore be invalid.

The remedy in this simple case is intuitively obvious. The experiment should be redesigned so that the order in which the types of text are used on each trial is randomized. It can be shown that when this is done, the ordinary techniques of analyzing variance become valid again, at least with good approximation. The principle of randomization admits of considerable elaboration, and together with appropriately formulated test and estimation procedures, makes available a variety of techniques for circumventing the difficulties caused by violation of the usual assumptions that the observation errors are statistically independent with equal variance.

As mentioned above, the F-test provides a means of determining whether or not there is significant variation from family to family. The simplest model of the family effect postulates that the family distributions differ only in their mean values. It is intuitively evident from random sampling considerations that increasing the number of experimental

observations should bring out differences between family means more clearly, and hence should improve the test in the sense that when there exist actual differences in the means, the null hypothesis (that there are no differences) should be rejected more strongly. This is indeed the case and furnishes a basis for deciding in advance how many observations should be taken on each family, i. e., for designing the experiment. For example, we may have no reason to expect one family mean to be any different from the others than another, and may wish to design the experiment so that if the largest difference between family means is Δ , the F-test will reject the null hypothesis with 90 percent probability. This probability is called the "power" of the F-test, and tables exist which enable one to determine what sample size should be used to achieve a given power with prescribed Δ . Figure 6.1 shows some calculations of the sample size required for various numbers of families. Here the number of samples per family is the same and Δ is expressed in units of σ , the standard deviation of each sample distribution.

Of course if the F-test rejects the null hypothesis, we will immediately be interested in estimating the magnitude of the family effect thus implied to exist. Techniques are available for this also, permitting us to establish intervals within which certain linear combinations of the family means are included with prescribed probability.

This brief sketch has attempted to convey some idea of the principal type of techniques available in the general area of analysis of variance. Considerable complexity is encountered when the data are classified into several groups of families simultaneously, both in the design of the experiment and the analysis of results. The underlying ideas are all basically similar, however, and represent extensions and elaborations of the elementary concepts discussed above.

No. of Families	Sample Size Required		
	$\Delta/\sigma = 1$	$\Delta/\sigma = 2$	$\Delta/\sigma = 3$
2	22-23	6-7	3-4
3	26	7-8	4-5
4	28-29	8	4-5
5	31-32	8-9	4-5

Figure 6.1 Sample Size vs. Number of Families
F-test - one-way Layout
Power = 90 o/o
Level of Significance = 0.05

[Data calculated from Pearson and Hartley
charts. See Scheffé, pp. 62-65]

6.2 Variations With Duration of Utterance

Turning now to the specific needs of our present program, let us consider first the variations in a speaker's library that may be attributed to fatigue and training effects.

Suppose each of a number J of speakers S_1, S_2, \dots, S_J is asked to read one sample of text in each of $I+1$ consecutive time intervals T_0, T_1, \dots, T_I . Let the first sample read by each speaker serve to furnish the standard library against which his I remaining sample libraries are compared by use of the index of similarity discussed above, and let y_{ij} be the index obtained when the j -th speaker reads in the i -th time interval. Our model then is

$$y_{ij} = \beta_i + e_{ij} \quad (i = 1, \dots, I; j = 1, \dots, J) \quad (6.1)$$

where β_i is the population mean of the samples in the i -th interval and the $\{e_{ij}\}$ are independent normal random variables with zero means and equal variances σ^2 . We wish to test the null hypothesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_I, \quad (6.2)$$

i. e., the hypothesis that there is no significant effect due to time of speaking.

The main consideration affecting the choice of text for each sample is that variability due to text differences may obscure the variation with time interval that we are interested in. One possibility is to use the same piece of text for every interval. This would involve word-for-word repetition by a given speaker several times, however, which is an unnatural mode of extended utterance. A better scheme would be to use several different passages of text X_1, X_2, \dots, X_I all of the same type and randomize the assignment of the different passages of text to speakers and time intervals so that systematic variation among the X 's will not upset our test.

We note from Figure 6.1 that 8 samples each from 4 families will permit an F-test at a level of significance of 0.05 that is 90 percent sure to reject H_0 if any pair of the population means $\{\beta_i\}$ differ by more than twice the standard deviation σ . The following scheme for assigning text material is thus a possibility:

		Speakers							
		S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
Time Intervals	T_1	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
	T_2	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1
	T_3	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2
	T_4	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3

Assignment of Text Material

With this arrangement each type of text appears once for each speaker and twice in each time interval. A similar layout with 3 time intervals, 3 types of text and 6 speakers will permit a test of nearly the same power, according to Figure 6.1.

The text sample X_0 used for the standardization run may be the same for each speaker, but should differ from the other X 's to avoid unnatural repetitions. The text samples themselves should be short enough to preclude smearing of the linguistic subsets within a single time interval, and yet long enough to include a representative selection of the most commonly occurring library references (see Sec. 2).

In order to test the null hypothesis (6.2), we observe first that for the model of (6.1) the following sum-of-squares decomposition may be made:

$$SS_T = SS_W + SS_B \quad (6.3)$$

$$SS_T = \sum_i \sum_j (y_{ij} - y_{..})^2, \quad \text{d. f.} = IJ - 1 \quad (6.4)$$

$$SS_W = \sum_i \sum_j (y_{ij} - y_{i.})^2, \quad \text{d. f.} = I(J - 1) \quad (6.5)$$

$$SS_B = J \sum_i (y_{i.} - y_{..})^2, \quad \text{d. f.} = I - 1 \quad (6.6)$$

Here a dot subscript means that the arithmetic average over the subscript normally in that position has been taken. If the random variables y_{ij} are independent and normal with identical variances, the sums of squares SS_T , SS_W and SS_B are independent random variables having chi-squared distributions with the degrees of freedom (d. f.) indicated. The statistic

$$F = \frac{I(J - 1)}{I - 1} \frac{SS_B}{SS_W} \quad (6.7)$$

then has an F-distribution with $\nu_1 = I - 1$, $\nu_2 = IJ - 1$ degrees of freedom. The test of the null hypothesis (6.2) consists of calculating the value of F from (6.7) and the observed data, choosing a level of significance α , looking up in a table of the F distribution the value F_α for which the cumulative F probability is $1 - \alpha$ and comparing the calculated F with F_α . If $F > F_\alpha$, we reject H_0 at the level of significance α and conclude that there is a significant effect due to time of speaking. If $F < F_\alpha$, we draw the opposite conclusion.

A word of reassurance concerning the assumptions upon which this test is based may be in order at this point. It is known that when the number of samples per family is equal for all families (which in our applications may

be made a requirement of the experimental design), then the effects of non-normality and inequality of variance on the F-test are very small for rather severe violations of these two assumptions. Non-independence is more serious, however, so that care must be taken in designing the experiment to eliminate undesired statistical dependence by randomization of the data and control of the conditions under which observations are taken.

Now let us suppose that in the above example we have rejected the null hypothesis, so that there is reason to believe that there are variations among the means β_i . We should now like to know something about the magnitudes of these implied variations. Here the following method* is particularly convenient. Let ψ denote any linear function of the β_i with constant coefficients c_i ,

$$\psi = \sum_i c_i \beta_i, \quad \left(\sum_i c_i = 0 \right). \quad (6.8)$$

then

$$\hat{\psi} = \sum_i c_i y_i. \quad (6.9)$$

and

$$\hat{\sigma}_{\psi}^2 = \frac{SS_W}{IJ(J-1)} \sum_i c_i^2 \quad (6.10)$$

are unbiased estimates of the mean and variances of ψ . It can be shown that any ψ of the form (6.8) satisfies the inequality

$$\hat{\psi} - S\hat{\sigma}_{\psi} \leq \psi \leq \hat{\psi} + S\hat{\sigma}_{\psi} \quad (6.11)$$

with probability $1 - \alpha$, where

$$S^2 = (I - 1) F_{\alpha}. \quad (6.12)$$

*See Scheffe, Sec. 3.4.

Here α and F_α are the same as used above in the F-test.

The flexible form (6.8) and the inequality (6.11) permit estimates of many quantities of interest, e. g., in the above example with appropriate choices of the $\{c_i\}$ we may estimate

$$\psi = \beta_4 - \frac{1}{3} (\beta_1 + \beta_2 + \beta_3)$$

$$\psi = \beta_4 - \beta_1 \tag{6.13}$$

$$\psi = (\beta_4 - \beta_3) - (\beta_2 - \beta_1)$$

and so on, and in each case if $\alpha = .05$ we can be 95 percent sure that these quantities fall within the corresponding intervals calculated from (6.11). In this way we can get an idea of how much the average library displacement increases with time, whether the increase is linear with time, etc., all of which add to our knowledge of the time variability of the sample space configuration.

We have illustrated in this section only one elementary method of studying time variability. Others exist, particularly methods that involve taking data under conditions that allow several factors in addition to time variability to be operative simultaneously (such as speaker and text variability) and analysis of results to isolate the effect of each factor singly. The added complexity of these methods may be justified if the more elementary ones prove unsatisfactory.

6.3 Variations Among Speakers

The one-way classification model discussed above may be used to study variations among speakers also. Here, however, the families would be speakers S_1, \dots, S_I and the samples from each would be obtained by having each read J test passages X_1, \dots, X_J , where the X 's now should be independent samples, preferably of a single type of text material. The times T_1, \dots, T_I at which the speakers read should be mixed up so that systematic variations with time will not obscure the speaker-to-speaker effect sought. Again a 4×8 layout will give a test which will reject with 90 percent certainty the null hypothesis that there are no speaker differences when actually there are two speakers who differ by $\Delta = 2\sigma$ or more. Accordingly, the following arrangement may be used:

		Text Samples							
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Speakers	S_1	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4
	S_2	T_2	T_3	T_4	T_1	T_2	T_3	T_4	T_1
	S_3	T_3	T_4	T_1	T_2	T_3	T_4	T_1	T_2
	S_4	T_4	T_1	T_2	T_3	T_4	T_1	T_2	T_3

Test Arrangement

The standard "calibration" library needed to calculate the interlibrary distances here would most appropriately be obtained for each X by a speaker S_0 different from the others. Each column in the above arrangement would then have its own standard against which its 4 sample libraries would be compared.

The F -test of the null hypothesis and the estimation of the magnitudes of the differences between speaker means in the event H_0 is rejected proceed exactly as before.

6.4 Variations Due to Type of Text

In order to determine whether the type of text significantly influences the library configuration, we may employ a similar one-way classification model, taking as families the text types X_1, \dots, X_I . A number of speakers S_1, \dots, S_J would read each type of text to obtain the necessary independent sequences of trials. Probably only two types of text need be tried: disconnected words and connected discourse. To obtain satisfactory power, according to Figure 6.1, about 6 speakers should be used, making the following arrangement appropriate:

		Speakers					
		S_1	S_2	S_3	S_4	S_5	S_6
Text Type	X_1	T_1	T_2	T_1	T_2	T_1	T_2
	X_2	T_2	T_1	T_2	T_1	T_2	T_1

Test Arrangement

Here the times of speaking T_1 and T_2 are distributed in such a way that the order in which the text material is read will not influence the variation due to type of material.

A suitable calibration scheme here would be to have each speaker read a standard passage X_0 to obtain a library against which his further utterances could be compared.

The test and estimation procedures already described apply here as well.

7. TECHNIQUES FOR FURTHER IMPROVEMENT OF COMPRESSION RATIO

7.1 Coding*

If a fixed set of library references can be found which gives an adequate representation of speech over a suitable range of speakers and text, we will be faced next with the problem of encoding the references for transmission. It can be shown that the process of encoding a source such as this for transmission through a prescribed noisy channel may, without loss of generality, be considered as a sequence of two operations: source encoding, or the transformation of the source output into a sequence of binary digits, and channel encoding, or the transformation of the binary sequences generated by the source encoder into suitable form for transmission through the noisy channel. The two encoders may be designed independently, the source encoder design depending only on the source characteristics and that of the channel encoder depending only on the channel characteristics. Accordingly, we shall consider here only the source encoding problem.

The choice of a "noiseless" coding scheme for unambiguous representation of a sequence of source references of "messages" in terms of binary digits depends on the relative probabilities of occurrence of the source references, their statistical independence, or lack of it, and whether the source rate is fixed or controllable.

If the source rate is controllable and the references occur independently, a lower bound exists for the average number of binary symbols per reference that can be used without the occurrence of ambiguity in decoding. The lower bound is the entropy of the message ensemble $H(R)$, which in this case is not greater than $\log_2 N$ bits, where N is the number of different references in the library. The lower bound may be approached to within one bit by making the length of the code word associated with each message

*See, for example, Fano, Transmission of Information, Wiley 1961.

as nearly equal as possible to (but not less than) \log_2 of the reciprocal probability of its occurrence. The lower bound may be approached even more closely by assigning code words to sequences of references or messages, rather than to individual references. Procedures for finding such optimum sets of code words are well known, and need not be discussed here.

The above scheme gets into trouble if the source generates messages at a fixed rate and the source encoder is required to transmit its output symbols at a constant rate, which is the situation we must deal with here. This occurs because the code words representing the messages have different lengths. It can be shown that any finite storage device designed to take up the difference between the rate at which messages are fed into the encoder and the rate at which they are transmitted from it will eventually overflow with probability 1. This difficulty can be overcome by using fixed-length code words for certain long sequences of messages, and allowing other long sequences with vanishingly small probabilities of occurrence to be confused in decoding. It can be shown that by encoding sufficiently long sequences in this way the lower bound mentioned above may be approached as closely as desired, and at the same time the probability of ambiguous decoding can be made as small as desired.

A more serious problem in our present application is the undoubted fact that successive references in a sequence representing speech are not independent. It can be shown that if the source is ergodic the theoretical situation here is formally the same as described above, with the exception that the entropy of the message ensemble, $H(R)$, which constituted the lower bound on average code word length above, is now replaced by a conditional entropy $H(A | A^\infty)$:

$$H(A | A^\infty) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n H(A_k | A^{k-1}). \quad (7.1)$$

Here $H(A_k | A^{k-1})$ is the conditional entropy of the k -th reference in a sequence given the $k-1$ preceding references, and is calculated from the corresponding conditional probabilities of particular references occurring in the k -th position for combinations of occurrences in the preceding $k-1$ positions. It is known that $H(A_k | A^{k-1})$ decreases or remains constant with increasing k , reflecting the fact that as more past history becomes known, the uncertainty about the next event decreases, on the average. The partial sums on the right in (3.24) also decrease or remain constant with increasing n , all of which is helpful in estimating the lower bound $H(A | A^\infty)$.

The number $H(A | A^\infty)$ is a characteristic of the source and is never larger than $H(R)$. When there is statistical dependence, therefore, and advantage can be taken of it, there exists the possibility of using shorter code words on the average for message representation, and consequently for achieving higher compression ratios. How much this advantage is and whether it is worth the trouble it would take to estimate it remains to be seen.

Turning now to what can be done practically, we see that the first and easiest thing to do is to get an estimate of $H(R)$, which may be regarded as a first approximation to $H(A | A^\infty)$. For this, the relative frequencies of occurrence of the various library references are needed. These may readily be tallied by the VDPS for a number of runs, working in several speakers and types of text to get a good sized representative sample of several minutes duration. To illustrate the simple calculations involved, we may take the 5 references found from the spherical clustering example of Section 5.2 and artificially consider them as a complete library. For their frequencies of occurrence we may add together the numbers of times each vector in each of the 5 clusters occurred and divide by the total. The table below shows the results.

<u>Reference</u>	<u>No. of Occurrences</u>	<u>P_j</u>	<u>-P_j log₂ P_j</u>	<u>Code Word</u>	<u>n_j</u>	<u>n_j P_j</u>
R ₁	1199	.402	.528	0	1	.402
R ₂	842	.282	.515	10	2	.564
R ₃	405	.136	.392	110	3	.408
R ₄	348	.116	.360	1111	4	.464
R ₅	<u>190</u>	<u>.064</u>	<u>.254</u>	1110	4	<u>.256</u>
TOTALS	2984	1.000	2.049 = H(R)			2.094 = \bar{n}

$$\log_2 5 = 2.322$$

We note that the entropy of the reference ensemble $H(R)$ is 2.049 bits per reference. If the references occurred with equal probability, the corresponding entropy would be $\log_2 5 = 2.322$, indicating that at least 0.273 bits per reference can be saved by optimum encoding. (Actually, ordinary encoding of the 5 references, which ignored their different rates of occurrence, would require code words of length 3 since non-integer lengths cannot be used.) Also shown is a set of optimum code words for reference sequences of length 1, for which the average code-word length \bar{n} is 2.094 binary symbols. This compares favorably with the value of $H(R) = 2.049$. The latter could be approached even more closely by encoding longer sequences of references. We observe that the most frequently occurring references have the shortest code words, and that any sequence of code words can be unambiguously decoded to retrieve the encoded reference sequence. In particular, it is apparent that the average data rate can be reduced by about 30 percent in this example by using optimum coding rather than ordinary 3-bit-per-reference coding.

In order to go beyond $H(R)$ in approximating $H(A | A^\infty)$ we would need further statistical information about the reference sequences, in particular information on conditional probabilities of the type

$$P(a_k | a_{k-1} \dots a_{k-n}),$$

in which a_k is the reference in an arbitrary position of a sequence and $a_{k-1} \dots a_{k-n}$ are the n preceding references. These a 's range over the entire library of references. Clearly, it would be a formidable and foolhardy undertaking to attempt the evaluation of these conditional probabilities to a high order. About all we could envision would be use of a high speed computer to tally the most frequently occurring sequences of length 2 or 3. In view of the monotonic property of the partial sums in (7.1), however, even one or two low-order conditional entropy terms would be of help in the estimation.

Actually, from a practical point of view we might argue that the lower bound $H(A | A^\infty)$ is chiefly of academic interest anyway. If the most we can hope to do is to get the frequencies of occurrence of sequences of length 2 or 3, why not encode these optimally and determine the average data rate experimentally? There may even be much simpler techniques such as run-length coding that will provide significant reductions of data rate. Everything depends on the data, and often a qualitative inspection will suggest ad hoc measures that are more effective and simpler to apply than more general theoretical approaches.

7.2 Segmentation

In the preceding sections of this report the efficient representation of human speech sounds has taken the form of finding a small library of instantaneous spectra one of which always matches the spectrum patterns of the talker with a satisfactory degree of fidelity. The purpose of investigations described in the preceding was to reduce the number of elements of the library and thus reduce the number of bits necessary for encoding and transmitting the sequence of spectra used in synthesizing speech.

Another and independent direction that speech bandwidth compression studies can take is directed at reducing the number of times a new spectrum pattern must be transmitted. Two different approaches can be taken to achieve this objective. One hinges on the statistical dependence of consecutive spectrum patterns, a

characteristic that can be exploited by coding techniques as discussed briefly in Section 7.1. A typical example of a coding technique is the application of "run-length" coding where the repetitive occurrence of identical spectra can be encoded by the transmission of the spectrum at the time of its first occurrence plus a numerical indication of the number of repetitions.

Changing the method of spectrum sampling by changing the sampling epoch is another approach to reducing the bit rate. This is particularly appealing if the receiver's primary interest is in the preservation of the information content of speech as, for instance, would be the case in voice controlled teletype applications. If speech could be segmented into the sequence of sounds that we wish to distinguish from one another, the number of spectrum samples could be reduced by preserving (say only) one sample per speech segment. Encoding and transmission would proceed in a manner similar to that described above under run-length coding.

The segmentation scheme studied briefly here is based on the fact that when significant speech events start, their energy distribution differs from that of the preceding sound. Change in energy distribution can be detected by measuring the sum of magnitudes of the vocoder channel time derivatives at every instant. The segmentation waveform, $s(t)$, is given below where a_i is the

$$s(t) = \sum_{i=1}^{18} \left| \frac{d a_i(t)}{dt} \right|$$

i -th channel amplitude in an 18-channel vocoder. This is illustrated in Figure 7.1 where each of the 18 channel amplitudes is represented by a number between 0 (not printed) and 7, and where $s(t)$ is printed below the digitized vocoder representation. The waveform $s(t)$ has the characteristic that it is small when the speech sound is quasi stationary--as during most extended vowel sounds--and is large at the boundaries of speech sounds. It will be noted that the above hypothesis concerning the nature of $s(t)$ is justified. Vertical lines above the sonagram and below the sequence of $s(t)$

samples denote the points in time where $s(t) > 9$ and where $s(t)$ has local peaks (plus to minus sign change in $\frac{d s(t)}{dt}$). Dotted lines indicate peaks where $s(t) > 7$.

The correlation between the location of speech segments marked off by the vertical lines and the speech sounds we wish to distinguish from one another is excellent. The resulting segmentation scheme never fails to partition the text where a partition is required, and only segments the text unnecessarily in a few instances.

If the instantaneous spectrum nearest the center of each speech segment is selected as typifying the sound represented by the segment, the selected spectra would form a good basis for the choice of sounds whose phonetic transcription is unambiguous. That is to say, the selected spectra represent sounds that are typical of the sound represented by the speech segment in which it is contained. Aside from its obvious application to speech transcription (conversion to a phonetically transcribed representation) these selected spectra (indicated with arrows in Figure 7.1 may be useful in bandwidth compression.

Suppose, for instance, that we generate new synthetic speech exclusively from the above described selected spectra. Between speech segment boundaries--as defined operationally by the simultaneous satisfaction of the inequality $s(t) > 9$ and the occurrence of a local peak of $s(t)$ --we reproduce repetitively the selected spectrum. The resulting new synthetic speech is shown in the digitized sonagram at the bottom of Figure 7.1.

The information content of this synthetic speech is much less, of course, than that of the original speech, and its transmission with reduced bandwidth lends itself readily to the application of run length coding. The 8370 bits of data necessary to transmit the sentence of Figure 7.1 can be compressed into only 1915 bits with the new synthetic speech. A factor of 4.36 bandwidth compression is thus achieved.

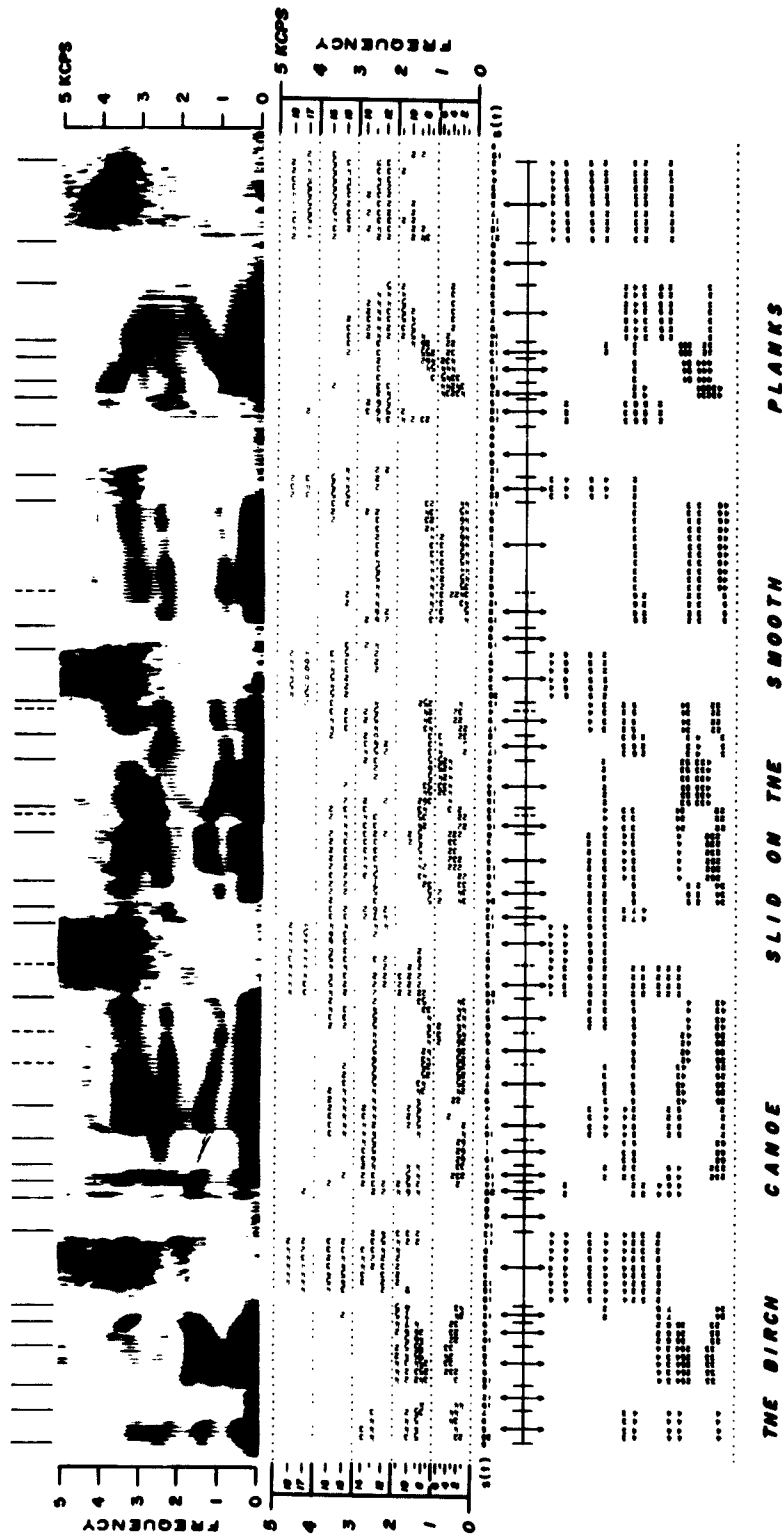


Figure 7.1

The resulting compressed speech was run through the AFCRL synthesizer and recorded on magnetic tape along with the original speech and the version appearing at the vocoder output, in order to facilitate quality comparisons (voicing and pitch information were artificially added). The compressed version was perfectly intelligible, and its quality was judged good.

8. RECOMMENDED PROGRAM OF EXPERIMENTATION

8.1 Determination of Equipment Errors

8.1.1 Data Needed:

a) Raw speech pattern sequences ($T_M = 0$) obtained by repeatedly playing the same 20 second recording of speech into the analyzer five times, with the memory cleared after each trial.

b) The speech pattern sequence ($T_M = 0$) obtained by playing a recording of the synthesizer output into the analyzer. The synthesizer input should be one of the 20 second library sequences obtained at the analyzer output in a).

8.1.2 Analysis:

a) Compute distances between every pair of library sequences obtained in 8.1.1a and average to find sampling error.

b) Order the sequence obtained in 8.1.1b with respect to the library sequence used as the synthesizer input, using the technique of Section 2.3.

Make 18 histograms, one for each channel, of the channel differences (with sign preserved) occurring in the pairs obtained by the above ordering. Examine these for skewness to judge quality of synthesizer adjustment.

8.2 Average Diameter and Spacing of Subsets

8.2.1 Data Needed:

Reference pattern sequence obtained with $T_M = 2$ for a 60 second segment of recorded speech.

8.2.2 Analysis

Apply the spherical clustering analysis of Section 5.1 to the first 20 second segment, the first 40 second segment and the entire 60 second segment of data obtained in 8.2.1. Use a clustering threshold of 4.

Read the new references obtained into the VDPS memory, and rerun the recording used in 8.2.1 with the machine in the classification mode. Record the synthesized speech and judge its quality. If it is satisfactory, repeat the clustering analysis with a higher threshold and synthesize again. When highest tolerable threshold is found, compute reference distance matrix to study subset configuration.

8.3 Rate of Subset Generation

8.3.1 Data Needed:

Number of references in library vs time for a 60 second segment of connected text, with machine thresholds of 2, 3, 4, and 5. Repeat for a second speaker.

8.3.2 Analysis:

Plot data as in Figure 4.1 and interpret as discussed in Section 4.

8.4 Variations With Duration of Utterance

8.4.1 Data Needed:

Reference libraries taken with $T_M = 4$ and 8 speakers, each reading 5 consecutive 20 sec. passages of text as discussed in Section 6.2.

8.4.2 Analysis:

Analysis of variance as discussed in Section 6.2.

8.5 Variations Among Speakers

8.5.1 Data Needed:

Reference libraries taken with $T_M = 4$ and 5 speakers, each reading 8 non-consecutive passages of text of 20 sec. duration, as discussed in Section 6.3.

8.5.2 Analysis:

Analysis of variance as discussed in Section 6.3.

8.6 Variations Due To Type of Text

8.6.1 Data Needed:

Reference libraries taken with $T_M = 4$ and 6 speakers each reading 3 passages of text of 20 second duration. Two of the text passages should be of different types.

8.6.2 Analysis:

Analysis of variance as discussed in Section 6.4.

8.7 Statistical Properties of Reference Sequences

8.7.1 Data Needed:

Any of the above data can be used if frequencies of occurrence are also tallied by the VDPS.

8.7.2 Analysis:

Determine average frequencies of occurrence and entropy of the reference ensemble. Use computer to tally digram and trigram frequencies if this appears to be justified. See discussion in Section 7.1.

8.8 Speech Segmentation

Further experimentation of the type described in Section 7.2 on longer utterances from a variety of speakers. In each case it is desirable to record the synthesized speech for comparative quality judgments.

BIBLIOGRAPHY

(1) Voice Data Processor:

Smith, C. P., A Method for Speech Data Processing By Means of a Digital Computer. ERD-CRRS-TM-58-103. AFCRC (1958).

Smith, C. P., An Approach to Speech Bandwidth Compression. Proceedings of Seminar on Speech Compression and Processing. AFCRC-TR-59-198. AFCRL (1959).

Smith, C. P., The Use of Digital Computers in Speech Analysis and Synthesis. AFCRL-TN-59-958. AFCRC (1959).

Final Report, Design and Development of a Digital Voice Data Processing System. AFCRL-62-314. Melpar, Inc. Contract AF19(604)-5579. (1962).

(2) Vocoder:

Dudley, H., The Carrier Nature of Speech. B.S.T.J. 19, pp 495-515. (October 1940).

Halsey, R. J. and Swaffield, J., Analysis-Synthesis Telephony with Special Reference to the Vocoder. J.I.E.E. 95 Part III, pp 391-411 (September 1948).

(3) Evaluation of Vocoder Performance as Function of Quantization:

Smith, C. P., Design vs. Performance Factors for Some Speech Compression Systems. ERD-CRRS-TM-61-9. AFCRL (November 1961).

Stevens, K. N., Hecker, M. H. L., and Kryter, K., An Evaluation of Some Speech Compression Systems. Report No. 914. Contract AF30(602)-2235. Bolt, Beranek, and Newman, Inc., Cambridge, Mass. (1962).

(4) Multiplex Equipment:

Hertz, D., Analog/Digital Multiplex Equipment for Voice Signal Processing. AFCRL-62-7. Contract AF19(604)-8042. Epsco, Inc., Cambridge, Mass. (1962).

(5) Amplitude Normalization Method.

U. S. Patent 2,901,697.

(6) Logic for Pattern Matching:

Straight, H. A. , Speech Data Processing in Real Time. AFCRL-62-719.
Contract AF19(604)-5579. Melpar, Inc., Falls Church, Va. (1962).